

Piotr Ladyzynski
Dorota G. Pijanowska
Adam Liebert *Editors*

Joint 20th Nordic-Baltic Conference on Biomedical Engineering & 24th Polish Conference on Biocybernetics and Biomedical Engineering

Joint Proceedings
of NBC 2025 and
PCBBE 2025, June
16–18, 2025,
Warsaw, Poland



Series Editor

Ratko Magjarević , *Faculty of Electrical Engineering and Computing, ZESOL, University of Zagreb, Zagreb, Croatia*

Associate Editors

Piotr Ładyżyński, *Warsaw, Poland*

Fatimah Ibrahim, *Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia*

Igor Lackovic, *Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia*

Emilio Sacristan Rock, *Electrical Engineering Department, Universidad Autonoma Metropolitana, Mexico DF, Mexico*

The IFMBE Proceedings Book Series is an official publication of *the International Federation for Medical and Biological Engineering* (IFMBE). The series gathers the proceedings of various international conferences, which are either organized or endorsed by the Federation. Books published in this series report on cutting-edge findings and provide an informative survey on the most challenging topics and advances in the fields of medicine, biology, clinical engineering, and biophysics.

The series aims at disseminating high quality scientific information, encouraging both basic and applied research, and promoting world-wide collaboration between researchers and practitioners in the field of Medical and Biological Engineering.

Topics include, but are not limited to:

- Diagnostic Imaging, Image Processing, Biomedical Signal Processing
- Modeling and Simulation, Biomechanics
- Biomaterials, Cellular and Tissue Engineering
- Information and Communication in Medicine, Telemedicine and e-Health
- Instrumentation and Clinical Engineering
- Surgery, Minimal Invasive Interventions, Endoscopy and Image Guided Therapy
- Audiology, Ophthalmology, Emergency and Dental Medicine Applications
- Radiology, Radiation Oncology and Biological Effects of Radiation
- Drug Delivery and Pharmaceutical Engineering
- Neuroengineering, and Artificial Intelligence in Healthcare

IFMBE proceedings are indexed by SCOPUS, EI Compendex, Japanese Science and Technology Agency (JST), SCImago. They are also submitted for consideration by WoS.

Proposals can be submitted by contacting the Springer responsible editor shown on the series webpage (see “Contacts”), or by getting in touch with the series editor Ratko Magjarevic.

Piotr Ladyzynski · Dorota G. Pijanowska ·
Adam Liebert
Editors

Joint 20th Nordic-Baltic Conference on Biomedical Engineering & 24th Polish Conference on Biocybernetics and Biomedical Engineering

Joint Proceedings of NBC 2025 and PCBBE
2025, June 16–18, 2025, Warsaw, Poland

Editors

Piotr Ladyzynski^{}
Nalecz Institute of Biocybernetics
and Biomedical Engineering, Polish
Academy of Sciences
Warsaw, Poland

Dorota G. Pijanowski^{}
Nalecz Institute of Biocybernetics
and Biomedical Engineering, Polish
Academy of Sciences
Warsaw, Poland

Adam Liebert^{}
Nalecz Institute of Biocybernetics
and Biomedical Engineering, Polish
Academy of Sciences
Warsaw, Poland

ISSN 1680-0737

ISSN 1433-9277 (electronic)

IFMBE Proceedings

ISBN 978-3-031-96537-1

ISBN 978-3-031-96538-8 (eBook)

<https://doi.org/10.1007/978-3-031-96538-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

This publication presents the Proceedings of the Joint 20th Nordic-Baltic Conference on Biomedical Engineering and the 24th Polish Conference on Biocybernetics and Biomedical Engineering (NBC 2025 & PCBBE 2025), held in Warsaw, Poland, on June 16–18, 2025.

This joint event represents a unique fusion of two long-standing traditions in biomedical science and technology, bringing together communities from the Nordic-Baltic countries, Poland, and the global biomedical engineering field. Held under the honorary patronage of Prof. Marek Konarzewski, President of the Polish Academy of Sciences, and organized under the common umbrella of the International Federation for Medical and Biological Engineering (IFMBE), the Committee of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences (CBBE PAS), and the Polish Society for Biomedical Engineering (PSBE), the event provided a vibrant forum for researchers, clinicians, industry representatives, and students to share the latest scientific achievements and discuss emerging trends in biomedical engineering.

The scientific program, developed under the guidance of an esteemed International Advisory Board and Scientific Program Committee, upheld the highest academic standards.

Full papers selected for presentation were peer-reviewed and are published in this volume of the IFMBE Proceedings, contributing to the ongoing dialogue and dissemination of knowledge within the global biomedical engineering community. Before acceptance for publication, each article underwent a single-blind peer review by at least two reviewers and went through a minimum of two review rounds. A total of 48 reviewers from 21 countries contributed to ensuring the high quality of the articles in this volume. The covered topics include biomedical signal and image processing, artificial intelligence in healthcare, biomaterials, nanotechnology, regenerative medicine, prosthetics, rehabilitation engineering, biomeasurements, biomedical sensors, and health technology assessment. Abstract submissions were also welcomed, ensuring a platform for both mature research and emerging ideas across virtually all areas of medical and biological engineering.

We would like to express our deep gratitude to all authors, reviewers, session chairs, and members of the scientific and organizing committees for their dedication and effort. The editors are also grateful to IFMBE, CBBE PAS, and PSBE for their continuous support in organization of the conference and the publication of the proceedings.

We believe that the contributions presented at NBC 2025 and PCBBE 2025 and published in this volume will serve as a valuable reference for researchers, practitioners,

and policymakers alike. Together, they represent a step forward in engineering the future of health.

Piotr Ladyzynski
Dorota G. Pijanowska
Adam Liebert

Organization

Honorary Patronage

Marek Konarzewski

President of the Polish Academy of Sciences

Organizers

Committee of Biocybernetics and Biomedical Engineering

Polish Academy of Sciences (CBBE PAS)

Polish Society for Biomedical Engineering (PSBE)

Nalecz Institute of Biocybernetics and Biomedical Engineering

Polish Academy of Sciences (IBBE PAS)

Conference Chairs

Adam Liebert

CBBE PAS, Poland

Marek Darowski

PSBE, Poland

Piotr Ladyzynski

IBBE PAS, Poland

International Advisory Committee

Adam Liebert (Chair), Poland

Martha Zequera (Co-chair), Colombia

Metin Akay, USA

Virginia Laura Ballarin, Argentina

Maria Fernanda Cabrera, Spain

Paulo de Carvalho, Portugal

Murilo Contó, Brasil

Marek Darowski, Poland

Yuri Dekhtyar, Latvia

James Goh, Singapore

Thordur Helgason, Island

Ashenafi Hussein Ababu, Ethiopia

Jari Hyttinen, Finland

Ernesto Iadanza, Italy

Håvard Kalvøy, Norway

Shankar Krishnan, USA
Igor Lacković, Croatia
Jaanus Lass, Estonia
Lenka Lhotska, Czech Republic
Kang-Ping Lin, Taiwan
Olof Lindahl, Sweden
Arūnas Lukoševičius, Lithuania
Ratko Magjarević, Croatia
Fabiola Marinez, Mexico
Leandro Pecchia, Italy
Subrata Saha, USA
Ichiro Sakuma, Japan
Elliot Vernet, Mexico
Hairong Zheng, China

Scientific Program Committee

Dorota G. Pijanowska (Chair), Poland
Diana Adliene, Lithuania
Milos Ajcevic, Italy
Hakan Aldskogius, Sweden
Iulian Antoniac, Romania
Piotr Augustyniak, Poland
Daiva Baltriukienė, Lithuania
Panagiotis Bamidis, Greece
Ryszard Białycki, Poland
Zhivko Bliznakov, Bulgaria
Hugh Byrne, Ireland
Marina Cabrin, Italy
Carmel Caruana, Malta
Frans Cornelissen, Netherlands
Gerard Cybulski, Poland
Todorka Dimitrova, Bulgaria
Arzum Erdem Gursan, Turkey
Ivo Fridolin, Estonia
Birgit Glasmacher, Germany
Joaquin Roca Gonzalez, Spain
Ludomira Granicka, Poland
Julius Griškevičius, Lithuania
Kārlis-Agris Gross, Latvia
Lejla Gurbeta Pokvic, Bosnia and Herzegovina
Marek Gzik, Poland
Ulrich Hofmann, Germany
Daoud Robert Iskander, Poland

Timo Jamsa, Finland
Tomaž Jarm, Slovenia
Janusz Jeżewski, Poland
Ákos Jobbágy, Hungary
Eliasz Kańtoch, Poland
Magdalena Kasprowicz, Poland
Alexei Katashev, Latvia
Anna Korzyńska, Poland
Małgorzata Kotulska, Poland
Michalis Koutinas, Cyprus
Elena Kozlova, Sweden
Gunta Krumina, Latvia
Jaanus Lass, Estonia
Dorota Lewińska, Poland
Maria Lindén, Sweden
Janis Ločs, Latvia
Roman Maniewski, Poland
Vaidotas Marozas, Lithuania
Jose Miguel Martin Martinez, Spain
Winfried Mayr, Austria
Mario Medvedec, Croatia
Kalju Meigas, Estonia
Robert Michnik, Poland
Valerie Moreno, France
Tomasz Moskalewicz, Poland
Bert Müller, Switzerland
Sergei Nazarenko, Estonia
Fredrik Nikolajeff, Sweden
Nicolas Pallikarakis, Greece
Tadeusz Pałko, Poland
Elżbieta Pamuła, Poland
Zbigniew Paszenda, Poland
Jolanta Pauk, Poland
Thomas Penzel, Germany
Ewa Piętka, Poland
Martins Piksis, Latvia
Kristjan Pilt, Estonia
Halina Podbielska, Poland
Ryszard Pohorecki, Poland
Dejan Popović, Serbia
Jevgenijs Proskurins, Latvia
Elizabeth Aune Ragnhild, Norway
Marius N. Roman, Romania
Anatoly B. Rozenfeld, Australia
Simo Särkkä, Finland

Heinrich Schima, Austria
Mário Forjaz Secca, Portugal
Karsten Seidl, Germany
Sudesh Sivarasu, South Africa
Janis Spigulis, Latvia
Johan Stiens, Belgium
Paweł Strumiłło, Poland
Fong-Chin Su, Taiwan
Andrzej Świerniak, Poland
Slavik Tabakov, UK
Sigitas Tamulevičius, Lithuania
Ewaryst Tkacz, Poland
Miroslav Trajanovic, Serbia
Milan Tyšler, Slovakia
Janusz Wróbel, Poland

Scientific Challenge Committee

Paulo de Carvalho (Chair)	University of Coimbra, Portugal
Piotr Foltynski	IBBE PAS, Poland
Jorge Henriques	University of Coimbra, Portugal
Piotr Ladyzynski	IBBE PAS, Poland

Conference Organizing Committee

Piotr Ladyzynski (Chair), IBBE PAS, Poland
Marta Baranowska
Hanna Bazior
Justyna Rogalska
Marek Rudnicki
Anna Stecka
Łukasz Wasilak
Barbara Wolff
Marcin Woźniak
Krzysztof Zieliński

Coorganizers



Endorsing Organization



Sponsoring Organization

The conference is co-financed from the state budget, allocated by the Minister of Science under the Excellent Science II Programme Contract No. KONF/SP/0624/2024/02



Contents

Biomedical Signal and Image Processing

Event-Related Potentials as Markers of Image Depth Perception.....	3
<i>Mehrdad Naderi, Albina Abdullayeva, Gunta Krumina, Reinis Alksnis, and Tatjana Pladere</i>	
Analysis of the Website Accessibility Influence on the Cognitive Workload of Users with Visual Impairments.....	10
<i>Piotr Zajac and Aleksandra Królak</i>	
Evaluation of Factors-of-Interest in Muscle Phantoms Based on Ultrasonic Data	25
<i>Vladislavs Agarkovs, Aleksandrs Sisojevs, Alexey Tatarinov, and Tamara Laimiņa</i>	
Nonlinear Analysis of ECG Signals in Arrhythmia Simulated by MPS450 Fluke and Measured with ADS1298 AFE.....	33
<i>Elzbieta Raus-Jarzabek and Elzbieta Olejarczyk</i>	
QRS Detection Based on the Matching Pursuit Algorithm.....	46
<i>Mohammed Owahidur Rahman, Piotr Augustyniak, and Elzbieta Olejarczyk</i>	
Influence of the 3D Printing Infill, Pattern and CT kV on the Hounsfield Units for Biomedical Applications.....	54
<i>Nikolay Dukov, Kristina Bliznakova, Yanka Baneva, and Zhivko Bliznakov</i>	
An Approach to Create a Computational Model of a Trabecular Bone.....	64
<i>Kristina Bliznakova, Dimitar Trizlov, Nikolay Dukov, Stoyan Ivanov, and Zhivko Bliznakov</i>	
Explainable Estimation of Blood Volume Pulse Signals from Video Sequences Using a Combination of Deep Learning Models and Signal Processing Methods	75
<i>Milena Sobotka, Kamil Kopryk, Muhammad Usman, and Jacek Rumiński</i>	

Artificial Intelligence in Healthcare

Custom Attention in Transformer Architectures for Multiclass ECG Beat Classification 93
Elia Pattacini and Vaidotas Marozas

Neurosymbolic AI Approach for Dry and Wet AMD Classification Using OCT Images 101
Aleksandar Miladinović, Alessandro Biscontin, Miloš Ajčević, Agostino Accardo, Dario Marangoni, Daniele Tognetto, and Leandro Infrerra

Early Prediction of ICU Readmissions Using Machine Learning in a Colombian University Hospital..... 111
Joel Andrés Lerma Pazos, J. E. Camacho Cogollo, and Jose Leonardo Mojica Peñaranda

Detection of Electronic Protected Health Information in DICOM Objects: Rule-Based Methods vs. LLMs..... 130
Dmytro Tkachenko, Milena Sobotka, Antoni Górecki, Jakub Kłopotek-Główczewski, Tomasz Neumann, and Jacek Rumiński

The Influence of Color on Semantic Ulcer Segmentation Using Deep Learning Models 147
Muhammad Usman and Jacek Rumiński

Estimation of Body Ratios from Low-Resolution, Dual-Energy X-Ray Images Using Convolutional Neural Networks..... 165
Kamil Janczyk, Jacek Rumiński, Tomasz Neumann, Aleksandra Szymczyk, and Piotr Wiśniewski

Artificial Intelligence Models for Age and Sex Estimation from Electrocardiograms 181
Daniel Francisco Albornoz, Cristian Ariel Sposato, Gustavo Javier Meschino, Héctor Santiago Manzollilo, and Virginia Laura Ballarin

Automatic Text Classification in Cardiac Risk Management: A Pilot Study. 193
Jacopo Vitale, Mark C. H. de Groot, Imo Höfer, Leandro Pecchia, Saskia Haitjema, and Bram Van Es

Genetic Algorithm for Predicting Primary Imatinib Resistance in Locally Advanced or Metastatic Gastrointestinal Stromal Tumors.....	200
<i>Martina Sassi, Jacopo Vitale, Maria Elisabetta Pagnano, Laura Risi Ambrogioni, Margherita A. G. Matarrese, Bruno Vincenzi, and Leandro Pecchia</i>	
Echocardiographic Image-Based Classification of Cardiac Amyloidosis: A Proof of Concept	211
<i>Maria Elisabetta Pagnano, Martina Sassi, Jacopo Vitale, Margherita A. G. Matarrese, Chiara Cirillo, Filippo Crispino, Giuseppe Limongelli, and Leandro Pecchia</i>	
Application of Artificial Intelligence in Psychiatric Neuroimaging: A Systematic Review	220
<i>Ilona Karpel, Ewelina Sobotnicka, and Ana Starcevic</i>	
Biomaterials, Nanomaterials, Regenerative Medicine, and Prosthetics	
Using Backscattered Electrons to Determine Surface Charge on Laser-Treated Titanium Alloys.....	247
<i>Arturs Abolins, Alberta Aversa, Marks Gorohovs, Yuri Dekhtyar, Maris Dortiņš, Lyubomir Lazov, Arturs Mamajevs, Hermanis Sorokins, and Edmunds Sprudzs</i>	
Effect of Titanium-Doped Hydroxyapatite on Biofilm Formation of <i>Pseudomonas aeruginosa</i>	258
<i>Ángel Rodrigo Ortiz Juárez, Jimena Mariette Robledo Dorantes, Araceli Zapatero-Gutiérrez, Josué García-Ávila, José Rafael Alanis-Gómez, and Fabiola Hernández-Rosas</i>	
Nanolayer of Monocrystalline Si Towards Nanodosimetry of Electron Radiation	267
<i>Galina Boka, Yuri Dekhtyar, Mirko Rocca, Elizabete Skrebele, and Hermanis Sorokins</i>	
Tunable Isotropic TPMS-Based 3D-Architected Scaffolds for Diabetic Bone Regeneration	278
<i>Fátima Morales-Gómez, Fabiola Hernández-Rosas, Araceli Zapatero-Gutiérrez, Erick Ramírez-Cedillo, and Josué García-Ávila</i>	
Biomedical Potential of the Methanolic Leaf Extract of <i>Verbesina persicifolia</i> : Antimicrobial Properties and Applications in Wound Healing ..	286
<i>Carlos Alberto López-Rosas, Seojin Han, Margarita V. Saavedra Vélez, José Rafael Alanis Gómez, and Fabiola Hernández-Rosas</i>	

Influence of Liner Material Poisson’s Ratio on Contact Pressure
Distribution in Transtibial Prosthetic Sockets: A Finite Element Analysis
Study 294
Agata Mrozek-Czajkowska

**Biomeasurements, Biomedical Sensors, Health Technology
Assessment, and Rehabilitation Engineering**

User Expectations on Wearable Medical Devices in Latvia..... 303
Nadezhda Semjonova, Roberts Malnieks, and Alexei Katashev

Subjective Evaluation of Phantoms in X-ray Breast Imaging Research:
Insights from Bulgaria 311
*Tsvetelina Teneva, Antonio Sarno, Nikolay Dukov,
Vasileios Eleftheriadis, Zhivko Bliznakov, Panagiotis Papadimitroulas,
and Kristina Bliznakova*

EGGS: Efficient Gathering and Structuring of Avian Egg Datasets 320
*Lukasz Roszkowiak, Pawel Pstrokonski, Wojciech Wojcik,
Krzysztof Damaziak, and Anna Korzynska*

Digital Measurement Modules for Non-invasive Monitoring
of Physiological Parameters Under Extreme Conditions 329
*Ewelina Sobotnicka, Jan Mocha, Grzegorz J. Nowak, Grzegorz Badura,
Aleksander Sobotnicki, and Łukasz Dziuda*

A Blood Pressure Normalization Method for Aortic Pulse Wave Velocity.. 343
Kristina Kõõts, Kristjan Pilt, and Ivo Fridolin

Intelligent System for Monitoring Sedentary Behaviour to Support Elderly
Care 352
*Santiago Lynett, Santiago Mesa, Wilder Castellanos, Martha Zequera,
and Pilar Castellanos*

A Robotic System Integrating the Semmes-Weinstein Monofilament
(SWM10g) for Automated Plantar Sensory Evaluation: Novel Approach
for Early DPN Detection..... 371
César Augusto Moreno Baracaldo and Martha Zequera Díaz

Enhancing Cognitive and Motor Skills Through Personalized Digital
Games 390
Petr Novak and Lenka Lhotska

Young Investigators Competition

Prognostic Value of Early Interictal Epileptiform EEG Patterns in Ischemic Stroke: A Predictive Model for Functional Outcomes.....	403
<i>Katerina Iskra, Edoardo Ricci, Andrea Bonini, Giovanni Furlanis, Michele Malesani, Paola Caruso, Marcello Naccarato, Paolo Manganotti, Agostino Accardo, and Miloš Ajčević</i>	
Effect of Huntington's Disease on Electrical and Mechanical Properties of Red Blood Cells in a Mouse Model.....	412
<i>Liliana Ponkratova, E Du, Hongyuan Xu, and Jianing Wei</i>	

Scientific Challenge Competition

Advanced Multi-class Wound Segmentation Using EfficientNet UNet: A Comprehensive Approach to Automated Wound Assessment.....	423
<i>Athanasios Kallipolitis, Melina Tziomaka, Argyrios Zafeiriou, and Ilias Maglogiannis</i>	
Towards Efficient Wound Management: An Automatic Deep Learning Model for Accurate Wound Image Segmentation.....	441
<i>Aleksandar Miladinović, Alessandro Biscontin, Andrea Bonini, Francesco Bassi, Simone Krešević, Alessandra Raffini, Katerina Iskra, Agostino Accardo, and Miloš Ajčević</i>	
Deep Learning-Based Wound Segmentation Using a U-Net Architecture....	449
<i>Martinus Boom</i>	
Dual-Attention U-Net++ with Class-Specific Ensembles and Bayesian Hyperparameter Optimization for Precise Wound and Scale Marker Segmentation	457
<i>Daniel Cieślak, Miriam Reca, Olena Onyshchenko, and Jacek Rumiński</i>	
Wound Area Detection with nnU-Net v2 Architecture.....	468
<i>Franciszek Kubala, Piotr Matusiewicz, Dominik Mika, Dominik Zawlocki, and Zbysław Tabor</i>	
Segmentation Challenge: Wounds and Markers in Focus.....	477
<i>Piotr Potrzebowski, Arkadiusz Karbowski, Kamil Sterniuk, Aleksandra Krajna, Zuzanna Cemka, Krystian Opała, and Jacek Rumiński</i>	
Automated Segmentation of Wound and Scale Markers Areas Using a U-Net-Based Neural Network.....	493
<i>Zuzanna Błaszowska, Karolina Dzwonkowska, Piotr Stefański, and Piotr M. Szczypiński</i>	

Unified Wound Detection and Segmentation Using YOLO: An Efficient
Approach for Accurate Wound Measurement..... 506
 *Mahdi Bazargani, Mohammad Javad Heidari, Mehrdad Anvari-Fard,
 and Hamid Soltanian-Zadeh*

Author Index..... 517

Biomedical Signal and Image Processing



Event-Related Potentials as Markers of Image Depth Perception

Mehrdad Naderi¹ , Albina Abdullayeva¹ , Gunta Krumina¹ ,
Reinis Alksnis² , and Tatjana Pladere¹

¹ Department of Optometry and Vision Science, Faculty of Science and Technology,
University of Latvia, Riga 1004, Latvia
gunta.krumina@lu.lv

² Statistical Research and Data Analysis Laboratory, Faculty of Science and Technology,
University of Latvia, Riga 1004, Latvia

Abstract. Stereoacuity, the ability to perceive depth based on binocular disparity, is a crucial function of the human visual system. Traditional methods for assessing stereoacuity rely on subjective patient responses, making them potentially prone to bias. This study used electroencephalography (EEG) to investigate event-related potentials (ERPs) as potential markers for the objective assessment of depth perception based on binocular disparity. EEG data were collected for 32 participants, categorized into normal and abnormal stereoacuity groups, under controlled viewing conditions. Comparing participants' brain activity with normal and abnormal stereoacuity, data analysis revealed significant differences in ERP amplitudes within the 75–185 ms time window following the onset of visual stimulus (related to P1 and P2 components of ERP), particularly in the parietal regions. These findings suggest that specific ERP components could serve as neurophysiological indicators for objectively assessing depth perception.

Keywords: Electroencephalography (EEG) · Event-Related Potential (ERP) · Depth Perception · Stereoacuity

1 Introduction

Stereoacuity, the ability of the visual system to perceive depth based on binocular disparity, is a critical aspect of human vision [1]. Accurately measuring stereoacuity is important for diagnosing and understanding binocular vision disorders. However, current clinical methods for assessing stereoacuity, such as the Titmus Fly Test and Randot Stereotest, rely on subjective patient responses. This dependence introduces variability and potential bias, compromising the reliability of results [2] especially in cases when there is a need to assess stereoacuity multiple times or regularly. Therefore, developing objective measures of depth perception is crucial for achieving more consistent and reliable assessments.

Electroencephalography (EEG) offers the potential to objectively evaluate depth perception by measuring brain activity in response to visual stimuli. Event-related potentials (ERPs), which are time-locked EEG responses to specific sensor or cognitive events, have been extensively used to explore various aspects of information processing [3]. Among ERP components, P1, P2, and P3 have been particularly studied in relation to visual perception and cognitive functions. These components hold promises as objective markers for assessing stereoacuity and depth perception, potentially providing a bias-free assessment of traditional subjective methods.

Recent studies have consistently emphasized the significance of the role of early ERP components in processing depth cues. The P1 component, typically observed around 100 ms post-stimulus, has been shown to exhibit sensitive to depth information, particularly stereoscopic depth cues [4]. The N2 component, commonly associated with conflict monitoring and cognitive control, has demonstrated heightened activity in response to depth-related incongruities, underscoring its role in resolving perceptual conflicts within three-dimensional spaces [5]. Moreover, the variations in the amplitude and latency of the P1 and P2 components can be used to distinguish between different levels of binocular disparity, reinforcing their potential as objective markers for assessing stereoacuity [6].

This study aimed to investigate the applicability of the ERP components (P1, P2, and P3) as objective markers of depth perception based on binocular disparity. By comparing these ERP components in individuals with and without normal stereoacuity, the research seeks to establish a reliable neurophysiological framework for the objective assessment of depth perception which is important in both clinical and research contexts.

2 Materials and Methods

2.1 Participants

The study involved 32 participants, divided into two groups: an abnormal stereoacuity group and a normal stereoacuity group. The abnormal stereoacuity group consisted of 16 participants (12 females and 4 males) with an average age of 23 ± 5 years. In this group, three participants had naturally deficient stereoacuity, and 13 participants had simulated stereoacuity deficiency induced by applying + 1.00 diopter sphere over non-dominant eye, reducing stereoacuity to worse than 100 arc seconds while maintaining binocular vision without suppression (assessed using the Worth 4 Dot test).

The normal stereoacuity group comprised 16 participants (12 females and 4 males) with an average age of 26 ± 6 years. All participants in this group had normal or corrected-to-normal visual acuity of 1.0 decimal units and stereoacuity of 40 arc seconds or better.

The study was approved by the Ethics Committee of the University of Latvia (ZD2019/20807) and conducted in accordance with the Declaration of Helsinki.

2.2 Visual Stimulus

The visual stimulus [2] consisted of four constant angular size rings with an outer diameter of 0.5° and a line width of 0.1° , positioned at a field eccentricity of 1.0° from the center of the planes of the volumetric display (Lightspace Technologies, model: X1406). The display contains 20 image depth planes coupled with a high-refresh-rate projection unit (60 Hz) [7]. The distance between two sequential image depth planes is 5 mm. The display has a resolution of 1024×768 pixels per image depth plane.

2.3 Set up

Participants were sitting at 90 cm viewing distance from the volumetric display. Twenty-one active electrodes were placed according to the international 10–20 system and the average of all electrodes was considered as reference (see Fig. 1).

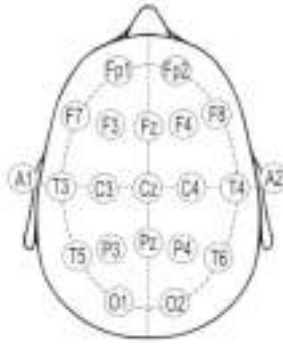


Fig. 1. The schematic of electrode placement on the scalp according to the international 10–20 system.

2.4 Procedure and Task

The experiment was conducted under scotopic conditions with an illuminance level of 1.2 lx. A total of 160 trials were presented, with the 3D effect appearing in 50% of the trials in a pseudo-randomized order. Each trial began with a fixation cross displayed at the center of the screen for 1 s, followed by the presenting of the rings. In the 3D trials, one ring (target) was displayed per one plane closer to the participant compared to three others. Participants were instructed to identify the target's position within the display and respond by selecting one of four directions (up, right, down, or left) using the arrow keys on a keyboard. In the 2D trials, all rings were shown on the same plane. For these trials, participants were instructed to press the spacebar on the keyboard responding that no target was identified.

2.5 Data Recording and Processing

The response time was recorded as the time starting from the initial demonstration of rings on the volumetric display till the arrow key button was pressed. The brain's electrical activity was recorded using a Nicolet v32 EEG system. Twenty-one active electrodes were positioned according to the international 10–20 system, with the average signal from all active electrodes used as the reference. EEG data were collected at a sampling rate of 1024 Hz, with a band-pass filter set between 0.1 and 70 Hz. Continuous recordings were performed throughout the trials, with electrode impedance maintained below 10 K Ω . Data files included triggers marking significant events, and EEG analysis was conducted using the EEGLAB 2023.1.0 toolbox integrated with MATLAB R2020a (MathWorks Inc., Natick, MA, USA).

For each participant, EEG data were cleaned of noise artifacts, including baseline drift, blinks, and muscle activity, using a two-step procedure that combined the software's built-in algorithm with manual variance inspection. The EEG signals were segmented into time-locked epochs of 1200 ms, aligned with the onset of each visual stimulus, and including a 200 ms peristimulus period. Epochs were then averaged based on image type (3D and 2D trials) across all participants.

The statistical analysis was performed in R 4.2.2. The Shapiro-Wilk normality test was used to verify the normal distribution of data. The effects of participants' group (normal stereoacuity vs. abnormal stereoacuity), image type (2D vs. 3D), and their interaction on response times were analyzed using a two-way repeated measures ANOVA. Raw EEG signals were analyzed to extract ERP components by segmenting the continuous signal into epochs for each channel in the parietal and occipital regions. The statistical analysis was performed for each time point within the -200 to 1000 ms window for each electrode in response to 3D images. A 95% confidence level was set for all analyses.

3 Results

Although individuals with abnormal stereoacuity appeared to have longer response times for both 3D and 2D images compared to those with normal stereoacuity (see Fig. 2), statistical analysis revealed no significant main effects for the participants' group ($F(1, 116) = 0.057$, $p = 0.81$, partial $\eta^2 = 0.0005$), image type ($F(1, 116) = 0.18$, $p = 0.67$, partial $\eta^2 = 0.0016$), and the interaction between these factors ($F(1, 116) = 0.701$, $p = 0.4$, partial $\eta^2 = 0.006$).

In contrast, significant differences in brain activity were observed between the normal and abnormal stereoacuity groups across specific time windows: 75–175 ms for the Pz electrode, 100–230 ms and 675–875 ms for the P3 electrode, 140–185 ms and 650–810 ms for the P4 electrode, 670–910 ms for the O1 electrode, and 270–310 ms and 645–975 ms for the O2 electrode. Figure 3 illustrates the topographic maps of the brain activity during early (75–185 ms) and late (645–875 ms) information processing.

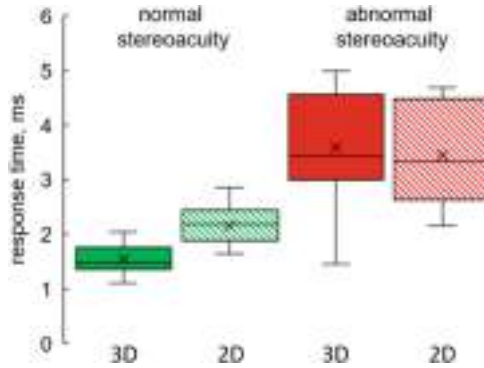


Fig. 2. Average response times to 3D and 2D images in individuals with normal (16 participants) and abnormal (16 participants) stereoacuity.

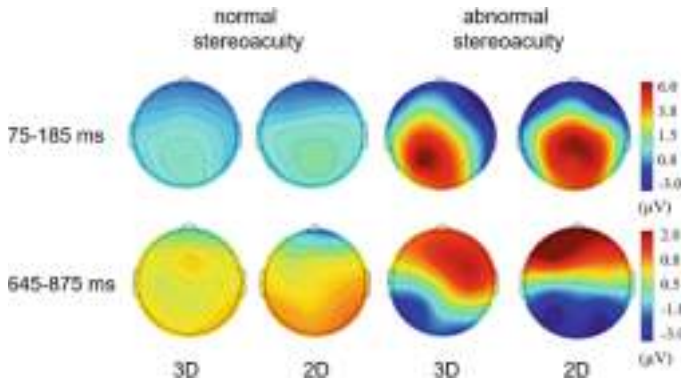


Fig. 3. Topographic maps of early (75–185 ms) and late (645–875 ms) ERP components in individuals with normal and abnormal stereoacuity.

4 Discussion

Traditionally, stereoacuity tests have been developed primarily relying on subjective patient responses, which can be influenced by individual variability and bias. This study aimed to explore the potential of ERP components as an objective alternative. While further research is needed to establish a standardized method for evaluating depth perception across various age groups and levels of depth perception deficiency, the current study represents the step forward toward this goal.

ERP components of EEG have been studied in various psychological [8, 9], neurological [10], and neuroscientific [11, 12] contexts. It is now well established that ERPs can serve as biomarkers for different neurophysiopathological conditions. The P1 component, the first positive deflection in the ERP waveform, typically appears around 100 ms post-stimulus [13]. It is primarily associated with early sensory processing and is often referred to as a sensory component. However, it has been suggested that P1 also plays a role in feedback from the higher visual cortex, attentional mechanisms,

and partially in perceptual processing [6, 14, 15]. In this study, higher amplitudes in the 75–185 ms time window have been observed in the parietal region (P3, P4, and Pz electrodes) in individuals with abnormal stereoacuity, thus, highlighting the relevance of the P1 and partially the P2 components in the assessment of depth perception based on binocular disparity. While differences in P1 and P2 amplitudes did not show statistical significance in the occipital region (O1 and O2 electrodes), visual inspection of the waveforms still indicated increased amplitudes. This might reflect additional brain processing in abnormal stereoacuity group allocated to focus attention on extracting visual information for further processing [15, 16]. Although simulated conditions were used in the abnormal stereoacuity group, it should be noted that four of 16 individuals had an abnormal stereoacuity that was not simulated. When their data were analyzed separately, similar results were observed. Nevertheless, future research should involve a larger number of individuals with stereoacuity deficiency to ensure the generalizability of the findings.

Revealed differences in the Late Positive Component (LPC), primarily linked to information integration, recognition, and memory processes (e.g., retrieval of detailed stimulus information), indicated stronger brain activity in individuals with abnormal stereoacuity. This may reflect the additional effort required for visual stimulus categorization [17, 18]. However, as the LPC is more associated with information integration than initial perception, it is not directly relevant to the objective of this study but could offer valuable insights for understanding how higher-order cognitive processes compensate for deficits in stereoacuity. Future research could explore this component further to elucidate its role in neural mechanisms underlying depth perception based on binocular disparity.

Depth perception deficiency may be related to factors such as asymmetrical refractive error [19], eye misalignment, congenital color vision deficiency [20], and amblyopia [21]. Of these conditions, amblyopia is the only abnormality associated with a neural system abnormality. Therefore, further studies are needed to evaluate brain activity in individuals with amblyopia while performing 3D visual tasks.

This study explored the potential clinical utility of ERPs in diagnosing binocular vision disorders, specifically, abnormal stereoacuity. Future research could focus on expanding the dataset to include participants of varying ages and a broader range of visual conditions.

5 Conclusion

This study aimed to evaluate the potential of ERP components as an objective biomarker for the assessing stereoacuity. Significantly higher amplitudes of early ERP responses, particularly the P1 and P2 components, were observed in the abnormal stereoacuity group compared to the normal stereoacuity group, predominantly within parietal regions. These findings align with the role of P1 and P2 in depth cue processing and the involvement of additional attentional mechanisms in individuals with abnormal stereoacuity. While the findings support the feasibility of using ERPs for objective evaluations, further research is necessary to generalize the results across diverse populations and develop the methodology for broader clinical application. In the future, this approach has the potential to

facilitate the assessment of depth perception, enabling more accurate diagnoses and interventions for binocular vision disorders.

References

1. Reichelt S, Häussler R, Fütterer G, Leister N (2010) Depth cues in human visual perception and their realization in 3D displays. *Proc SPIE* 7690:76900B
2. Naderi M, Pladere T, Alksnis R, Krumina G (2023) Brain activity underlying visual search in depth when viewing volumetric multiplanar images. *Sci Rep* 13:7672
3. Lopez-Calderon J, Luck SJ (2014) ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci* 8:213
4. Liu B, Meng X, Wu G, Dang J (2013) Correlation between three-dimensional visual depth and N2 component: evidence from event-related potential study. *Neuroscience* 237:161–169
5. Pegna AJ, Darque A, Roberts MV, Leek EC (2018) Effects of stereoscopic disparity on early ERP components during classification of three-dimensional objects. *Q J Exp Psychol* 71(6):1419–1430
6. Omoto S et al (2010) P1 and P2 components of human visual evoked potentials are modulated by depth perception of 3-dimensional images. *Clin Neurophysiol* 121(3):386–391
7. Osmanis K et al (2018) Advanced multiplanar volumetric 3D display. *Proc SPIE* 10555:116–128
8. Treder MS, Blankertz B (2010) (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behav Brain Funct* 6:28
9. Agam Y, Sekuler R (2007) Interactions between working memory and visual perception: An ERP/EEG study. *Neuroimage* 36(3):933–942
10. Nidal K, Malik AS (2014) EEG/ERP analysis: methods and applications. CRC Press, USA
11. Philips S, Takeda Y (2009) An EEG/ERP study of efficient versus inefficient visual search. In: *Proceedings of the annual meeting of the cognitive science society*, vol 31, no 31, pp 383–388
12. Bieniek MM, Frei LS, Rousselet GA (2013) Early ERPs to faces: aging, luminance, and individual differences. *Front Psychol* 4:268
13. Schindler S, Bruchmann M, Gathmann B, Moeck R, Straube T (2021) Effects of low-level visual information and perceptual load on P1 and N170 responses to emotional expressions. *Cortex* 136:14–27
14. Taylor MJ (2002) Non-spatial attentional effects on P1. *Clin Neurophysiol* 113(12):1903–1908
15. Fu S, Fedota JR, Greenwood PM, Parasuraman R (2010) Dissociation of visual C1 and P1 components as a function of attentional load: An event-related potential study. *Biol Psychol* 85(1):171–178
16. Luck SJ, Heinze HJ, Mangun GR, Hillyard SA (1990) Visual event-related potentials index focused attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. *Electroencephalogr Clin Neurophysiol* 75(6):528–542
17. Sun J, Osth AF, Feuerriegel D (2024) The late positive event-related potential component is time locked to the decision in recognition memory tasks. *Cortex* 176:194–208
18. Desauvay P et al (2017) Impact of semantic relatedness on associative memory: An ERP study. *Front Hum Neurosci* 11:335
19. Joo HJ, Choi DG (2023) Analysis of stereopsis and fusion in school-aged children with reduced visual acuity due to refractive error. *PLoS ONE* 18(4):e0284112
20. Ozates S, Sekeroglu MA, Ilhan C, Doguizi S, Yilmazbas P (2019) Depth perception in patients with congenital color vision deficiency. *Eye* 33(4):674–678
21. Webber AL, Schmid KL, Baldwin AS, Hess RF (2020) Suppression rather than visual acuity loss limits stereoacuity in amblyopia. *Invest Ophthalmol Vis Sci* 61(6):50



Analysis of the Website Accessibility Influence on the Cognitive Workload of Users with Visual Impairments

Piotr Zajac^(✉) and Aleksandra Królak

Institute of Electronics, Lodz University of Technology, Łódź, Poland
piotr.zajac@dokt.p.lodz.pl, aleksandra.krolak@p.lodz.pl

Abstract. The aim of this study was to analyze the relationship between the cognitive load when using selected online services and the level of accessibility of these services. The Zara website was selected for the study as a website with a low level of accessibility and the Zalando website as a website with a good level of accessibility. The accessibility of websites was determined on the basis of their compliance with the web content accessibility guidelines (WCAG) using automatic online tools supported by an expert evaluation. The cognitive workload was estimated based on the analysis of the time-domain HRV (heart rate variability) parameters and NASA-TLX inspired questionnaires. In the study participated 5 persons with visual impairments, 4 men and 1 woman, on average 40 years old. The results indicate a significant influence of the level of website accessibility on the mental workload of the users with visual impairments.

Keywords: cognitive workload · web accessibility · ECG · HRV · self-assessment questionnaire

1 Introduction

Digital accessibility means creating websites, mobile applications, multimedia, documents, and electronic services in a universal way, ensuring that the widest range of users can access them. According to the World Health Organization (WHO), at least 2.2 billion people have vision impairments, and an estimated 1.3 billion people experience significant disabilities, constituting 16% of the global population [4,5].

Cognitive workload (CWL) refers to the mental effort required to process information and complete a task, encompassing the demands placed on an individual's cognitive resources [1,6]. It is influenced by factors such as task complexity, time pressure, and individual capacity. An excessive amount of information leads to working memory overload and an increase in CWL. This results in the loss of conveyed information, problems with concentration, reduced efficiency, and heightened stress. In 1988, John Sweller developed the theory of CWL [3]

and suggested that our working memory is able to hold only a small amount of information at any one time. Studying CWL can have an impact on the design of the online services and their interfaces by determining the right balance between the amount of information for users and the optimal level of usability and accessibility of these services. It is beneficial to use known schemes and recognizable elements for users. Effective CWL management can make a service more user-friendly, increase user satisfaction, and consequently increase its popularity [1, 2].

Subjective measures of CWL rely on self-reported assessments, typically obtained through rating scales. Common tools include the NASA Task Load Index (NASA-TLX) [7], which evaluates workload based on dimensions like mental demand, temporal demand, and effort, and the Borg Scale [8], which measures perceived exertion. These scales provide valuable insights by capturing the individual's subjective experience, complementing objective physiological and performance-based measures. CWL estimation through biosignal analysis involves interpreting physiological responses to assess mental effort [9]. Key biosignals used include heart rate variability (HRV) [10], electroencephalography (EEG) [11], functional Magnetic Resonance Imaging (fMRI) [12], functional Near-Infrared Spectroscopy (fNIRS) [13], respiratory activity [14], and eye activity [10].

HRV reflects autonomic nervous system dynamics, with reduced variability often indicating heightened cognitive load [15]. EEG measures brain activity, providing insights into workload through specific frequency bands, such as increased theta activity and suppressed alpha waves during higher mental effort [16]. CWL can also be detected through breathing rate monitoring by analyzing changes in respiratory patterns, which are influenced by the autonomic nervous system (ANS). When cognitive demands increase, the sympathetic nervous system activates, often resulting in faster and shallower breathing, reduced respiratory variability, and changes in the synchronization between breathing and heart rate [17].

These methods enable real-time, non-invasive assessment of workload in contexts like task performance, human-machine interaction, and training environments. Website accessibility and usability significantly influence CWL by determining how easily users can interact with online content. Accessibility ensures that websites are navigable and functional for all users, including those with disabilities, by adhering to guidelines such as the Web Content Accessibility Guidelines (WCAG). Poor accessibility, such as insufficient contrast, unclear navigation, or lack of alternative text for images, forces users to expend additional cognitive resources to process and interpret information. Similarly, usability focuses on the efficiency and intuitiveness of a website's design. Complex layouts, inconsistent interfaces, or excessive information can overwhelm users, increasing their cognitive load. Conversely, well-designed websites with clear navigation, logical organization, and responsive elements reduce mental effort, enabling users to focus on their tasks without unnecessary distractions.

The relationship between CWL and interface design has been explored in several studies. In the research by Albers [18] a series of tasks on two websites

were given to the group of participants to verify if the task difficulty increases the cognitive workload measured by a tapping test. The influence of the website usability on cognitive load was analyzed in [19] when using NASA TLX, Sternberg Memory Test and the tapping test for CWL estimation. Rating scale NASA TLX was also used in [20] to check the influence of a website contrast on CWL, however in this research no impact of the contrast level on the scores was found. The GSR signal was used in [21] to detect frustration of visually impaired and sighted users while performing tasks on the websites using smartphones. The mobile and PC versions of the interfaces and their impact on CWL were analyzed using a low-cost EEG device in [22]. The results indicated that the website design significantly influences the level of the CWL of the user.

Research indicates that complex web interfaces can increase cognitive load, particularly for individuals with cognitive disabilities. However, there are very few studies employing physiological measures for CWL estimation to evaluate the impact of inaccessible and unusable website design on the mental effort and stress level of the users.

The aim of this study was to verify whether the level of web accessibility implemented in the selected websites influences the CWL of users with vision impairments, as measured via heart rate variability analysis.

2 Materials and Methods

2.1 Web Accessibility

The most recognized standard for digital accessibility is Web Content Accessibility Guidelines (WCAG) 2.1 developed by the World Wide Web Consortium (W3C) [26]. Testing the digital accessibility of websites can be approached in various ways. The easiest solution to use are tools for automatic accessibility testing. There are many such solutions, simple and more advanced. However, they cannot verify all the guidelines. A more advanced method is expert evaluation. It involves a comprehensive check of a given portal. In this case, it is possible to use e.g., Website Accessibility Conformance Evaluation Methodology [27]. Next possibility is the research study with the participation of disabled people (user testing). The study can take various forms. Often, a task scenario is developed, followed by a questionnaire to collect assessments. These can be both quantitative and qualitative assessments. The best results are achieved by combining these methods, but it is time consuming.

In the presented study, automatic methods were used for website accessibility evaluation and supported by an expert evaluation.

2.2 Cognitive Workload Assessment with HRV

Analysis of heart rate variability (HRV) in the time domain provides important parameters that describe heart rate fluctuations over time. These methods are useful for assessing autonomic function and can help to better understand the regulation of the autonomic nervous system on the heart [23]:

- NN Medium Range: Measures the average interval between successive beats of the heart. The overall HRV, represented by mean NN intervals, decreases with the increase of the cognitive load [25].
- SDNN (Standard Deviation of NN Intervals): Standard deviation of the NN intervals reflects all the cyclic components responsible for variability over the entire recording period. The decrease in the value of SDNN is correlated with the increase of the mental workload [23, 24].
- RMSSD (Root Mean Square of Successive Differences): The square root of the average of square differences between successive NN intervals, useful to assess the short-term variability attributed to parasympathetic activity. Similarly to SDNN, RMSSD value decreases with the increase of the CWL [23, 24].
- NN50 and pNN50: Count of pairs of successive NN intervals that differ by more than 50 milliseconds; pNN50 is the ratio of NN50 to the total number of NN intervals. It has been found that the values of NN50 and pNN50 decrease with the increase of the task difficulty [25].
- Heart Rate (HR) is the number of times the heart beats within a certain time period, usually a minute. Increased cognitive workload leads to the increase in heart rate [24]

In the presented study, all time-domain HRV parameters described above were analyzed.

2.3 Measurement Procedure

Web Accessibility Analysis. The entire study and website accessibility evaluation were conducted on the same device. This device was a laptop running Windows 10 22H2 64-bit, with Mozilla Firefox 125.0.3 as the browser and NVDA 2024.1 as the screen reader. Two tools, each offering a different approach, were used for the accessibility evaluation. The first tool was the Markup Validation Service created by the W3C consortium, sometimes referred to as the “Nu Html Checker”. This tool enables checking HTML compliance with the standard, allowing for the detection of errors and warnings. It also checks the correctness of CSS styles. However, the results need to be appropriately interpreted, as some detected errors may not actually be errors. The second tool is the WAVE Web Accessibility Evaluation Tools browser extension. WAVE can identify numerous accessibility and Web Content Accessibility Guidelines (WCAG) errors, but also facilitates human evaluation of web content. Both approaches have been used.

ECG Signal Acquisition. The electrocardiogram (ECG) was recorded using the BIOPAC system. This system is widely used for recording and analyzing physiological signals, including ECG, and is suitable for research in psychophysiology, neuroscience, biology, and other fields. For the ECG signal acquisition, one channel of the MP36 module of the BIOPAC system was used. The signal was registered in the frequency range of 0.5–100 Hz according to AHA requirements, with a sampling frequency of 200 Hz. The ECG signal was recorded using Einthoven’s lead. The study involved five participants with visual impairments,

all of whom held official disability certificates. The group consisted of four men and one woman, with an average age of 40 years. Among them, four participants were completely blind, while one had minimal vision, limited to light perception. All participants were experienced screen reader users, relying on assistive technologies for daily computer interactions. Additionally, they were familiar with online shopping, using both websites and mobile applications to navigate digital marketplaces independently.

Scenario. The testing scenario is shown in Table 1. It involved the completion of nine tasks. These tasks were intended to be performed using the NVDA screen reader and a keyboard.

Table 1. Website accessibility testing scenario

ID	Task
1.	Rest time (2–5 min)
2.	Free exploration of the website (1 min)
3.	Find “PRIVACY POLICY”, open it in PDF and see if you can read it
4.	Return to the home page. Use the search engine to find a “winter jacket”. Try adding any jacket in size ‘L’ to your cart (shopping bag)
5.	Go to the men’s T-shirts section. Change the price range in the filters. Choose any T-shirt. After going to the T-shirt subpage, try to choose the color and size and add it to the cart
6.	Please see the measurement chart for any T-shirts. Please check the measurements for XL
7.	Go to cart. Change the number of T-shirt to 3 and proceed (to purchase)
8.	Register a new account. Fill in the details and check the boxes
9.	Return to the home page. Try to find company details or contact details

After completing the measurements, participants were given a self-assessment questionnaire. It consisted of two parts. In the first part, participants were asked to rank the tasks from easiest (1) to most difficult (8), without rest time (task 1). In the second part, they rated the subjective evaluation scale for mental workload (1 – very easy, 10 – unable to complete), without rest time (task 1). The questionnaire was inspired by the NASA Task Load Index (NASA-TLX), but the full version was not used in this study.

2.4 Signal Processing

Signal processing algorithms were implemented in the Python programming language, utilizing classical libraries such as Numpy, Matplotlib, Pandas, and Scipy, as well as more advanced libraries for ECG and HRV signal analysis, such as pyHRV and Biosppy. The following methods were used:

- Bandpass Filter: 0.5–45 Hz bandpass filter. The lower limit of 0.5 Hz helps to eliminate low-frequency noises such as patient movement or breathing that can distort the ECG signal, while the upper limit of 40 Hz reduces high frequency noise such as that resulting from electromagnetic interference or from electronic devices in the surrounding area. Furthermore, the significant complexes of the signal are within the pass band, so it is possible to improve the quality without distorting the signal.
- Notch filter: remove 50 Hz noise (AC filter).
- Manual removal of artifacts or incorrectly determined R peaks.
- Detection of R-peaks: the R-peak is part of the QRS complex, which is essential for subsequent heart rate variability (HRV) analysis.
- R-peaks Analysis: Once the signal segments of interest and the corresponding R peaks were obtained, the time intervals corresponding to the R peaks in the two segments were calculated; they indicate at what instant in milliseconds an R peak occurred. Finally, the NN intervals were calculated, i.e., the time intervals between one R peak and its next (in ms).
- Time domain parameters: In the time domain 7 parameters were calculated: SDNN, SDANN, RMSDD, NN50, PNN50, mean HR and mean NNI difference.
- The paired t-test was applied to the obtained results in order to compare the means and to determine which parameters are statistically significant. The difference between two sets was considered statistically significant if the p-value from the paired t-test was less than the chosen significance level equal to 0.05.

The methodology previously discussed has led to the evaluation of several HRV parameters: mean value of NNI, SDNN, RMSSD, NN50, pNN50 and average Heart Rate (HR). The parameters were calculated in the first step for three signal acquisition sections: rest, considered as baseline, during performing scenario tasks on Zara website and while performing tasks on Zalando website. The ECG signals collected from 4 out of 5 participants were analyzed. One participant was excluded from the analysis because his ECG signal showed a significant arrhythmia, which could influence the HRV signal.

3 Results

3.1 Web Accessibility Analysis

The evaluation of the Zara website using the Markup Validation Service revealed 288 errors and 89 warnings. The main issues identified were:

- Duplicate, bad, and empty attributes (e.g., ID)
- Incorrect use of the Accessible Rich Internet Applications Suite (WAI-ARIA)
- Incorrect creation of lists ()
- Missing alt attributes for images, which are required.

Combining automated accessibility checking and expert evaluation allowed for the identification of additional accessibility issues. The most significant problems identified were:

- Color Contrast: Insufficient color contrast between text and background, making it difficult for users with visual impairments to read content. WCAG 1.4.3 requires a contrast ratio of 4.5:1.
- Keyboard Navigation: Inconsistencies and difficulties in navigating the website using only a keyboard. The homepage contains numerous elements that are not visible but are read multiple times, particularly information about social media. Additionally, content appears and disappears in an unusual manner. The site is challenging to navigate using a keyboard. (WCAG 2.1.1)
- Focus is hard to see, which violates WCAG 2.4.7.
- Screen Reader Compatibility: Inadequate labeling and descriptions for screen readers, leading to a poor user experience for visually impaired users.
- A mechanism is not available to bypass blocks of content that are repeated on multiple Web pages. (WCAG 2.4.1)
- There should be an option to stop all animations on the page (WCAG 2.2.2)

However, the biggest accessibility issue with the Zara website is the overlay that was supposed to ensure accessibility. This solution is highly problematic, as it introduces unnecessary features, redundant keyboard shortcuts, and often does not function correctly. It is much better to provide the website with basic functionalities rather than paying extra for ineffective overlays. For the Zalando website, the Markup Validation Service detected 116 errors and 40 warnings, most of which are minor attribute issues. The validator also identified the absence of the required alt attribute for images and the lack of a first-level heading. An analysis of the website using the WAVE tool indicates the need to improve the page's heading hierarchy (WCAG 2.4.6). Additionally, a form control does not have a corresponding label. The site also contains several empty buttons. Despite a few errors, the Zalando website is carefully designed and is friendly for screen reader users.

3.2 Cognitive Workload

In this section the results of HRV signals analysis are presented. The HRV signals were calculated from single lead ECG. Examples of the recorded ECG signals are presented in Fig. 1.

The computed values of the HRV parameters are presented in Fig. 2. It can be observed that for all participants, the mean HR increased while performing tasks on Zara website comparing them with baseline values as well as to Zalando values. It can be also seen that the rest of the parameters show the decrease of the values for Zara website tasks in comparison to the HRV parameters values computed while the tasks on Zalando webpage were executed.

The mean values $\pm$ SD of selected HRV parameters are presented in Fig. 3. These average results also show the dependence between the increased cognitive

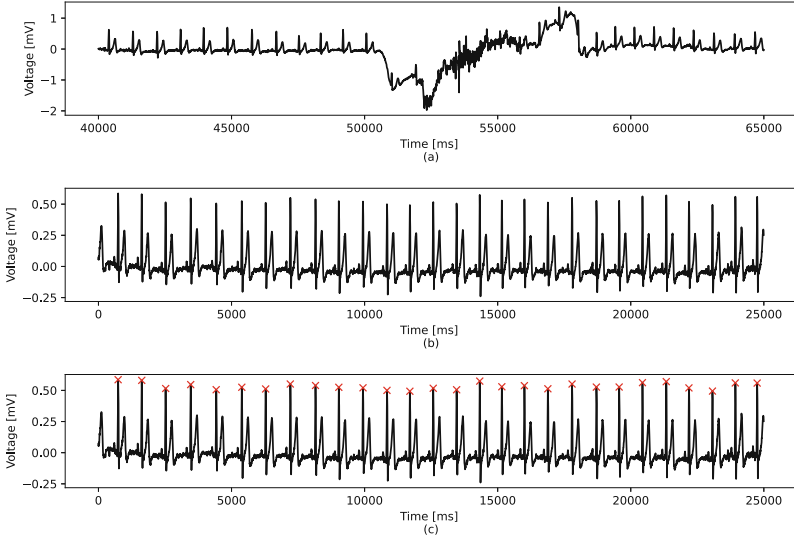


Fig. 1. Sample ECG signals (a) with artifacts, (b) normal ECG, (c) normal ECG signal with detected R peaks.

workload while working on Zara website with changes in HRV parameters. It can be observed that the average values of SDNN, RMSSD, NN50 and PNN50 decrease with the assumed difficulty of tasks and related CWL.

The second part of the analysis included computation of the selected HRV parameters for 5 tasks (Task 3–7). Tasks 8 and 9 were excluded from the analysis, as not all participants completed them. A general trend can be observed that all the parameters show increased cognitive load when performing tasks on Zara website, comparing to Zalando website (Fig. 4).

Despite the low number of participants, the statistical analysis using the paired t-test was performed. The p-values are presented in Tables 2 and 3. The statistically significant differences were found for four out of six HRV parameters when comparing their values calculated for completing the scenarios on Zalando and Zara websites. It implies that there is a meaningful difference between the level of CWL while performing designed tasks using these two online services. None of the HRV parameters showed statistically significant difference in case of comparing data for the rest period and performing tasks on Zalando website. It suggests that this online service does not cause notable increase in CWL. In case of the comparison of resting period and scenario completion on Zara website, three HRV parameters showed significant statistical differences, namely NNI, PNN50 and mean value of HR.

Statistical analysis performed for each scenario task separately showed significant statistical differences in case of each task, but for various HRV parameters. The smallest statistical difference was obtained for Task 4 between the Zara and Zalando tasks regarding the assessed HRV parameters.

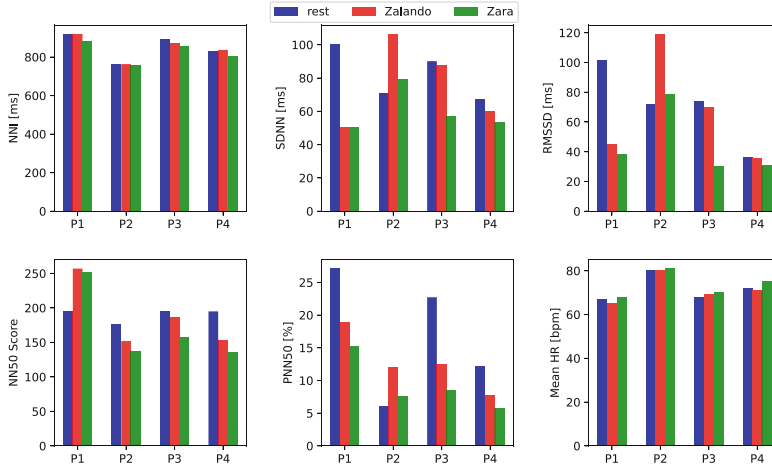


Fig. 2. HRV parameters values comparison for resting state, Zara tasks and Zalando tasks for each participant ($N = 4$).

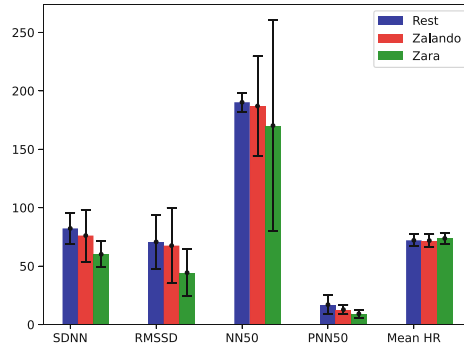


Fig. 3. Mean $\pm$ SD values of selected HRV parameters values for resting state, Zara tasks and Zalando tasks, calculated for 4 study participants ($N = 4$).

Table 2. The p-values calculated using paired t-test comparing resting period and scenario completion with Zara and Zalando websites for 4 study participants ($N = 4$)

HRV parameter	Zara vs. Zalando	Rest vs. Zara	Rest vs. Zalando
NNI	0.043	0.043	0.706
SDNN	0.084	0.175	0.749
RMSSD	0.051	0.205	0.890
NN50	0.038	0.495	0.896
PNN50	0.008	0.043	0.327
Mean HR	0.048	0.035	0.495

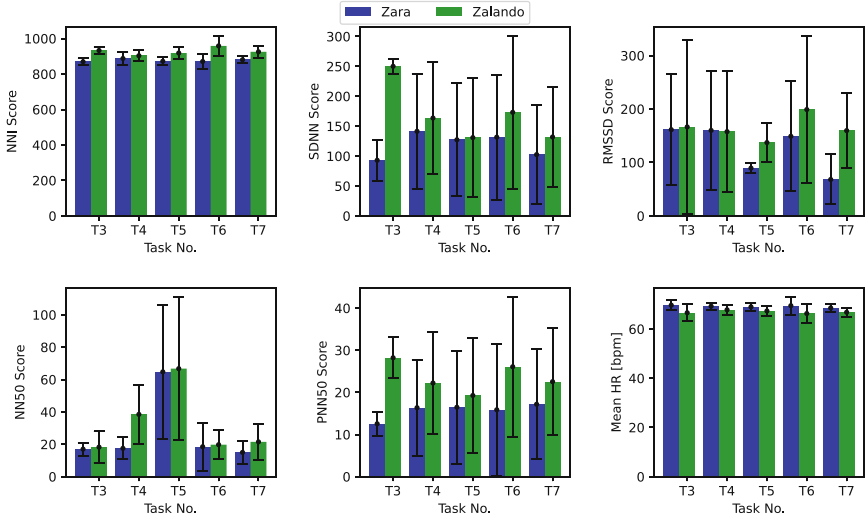


Fig. 4. Mean $\pm$ SD values of selected HRV parameters values for Zara tasks and Zalando tasks separately ($N = 4$).

Table 3. The p-values calculated using paired t-test for each performed scenario task for 4 participants ($N = 4$)

HRV parameter	Task 3	Task 4	Task 5	Task 6	Task 7
NNI	0.031	0.342	0.008	0.015	0.091
SDNN	0.002	0.031	0.371	0.474	0.023
RMSSD	0.959	0.492	0.040	0.413	0.006
NN50	0.730	0.040	0.908	0.863	0.373
PNN50	0.013	0.035	0.195	0.070	0.011
Mean HR	0.029	0.203	0.038	0.022	0.151

3.3 Self-assessment Questionnaire

The results of the questionnaires completed by the study participants are presented in Tables 4 and 5. These tables also include the average scores for each task. Table 4 shows the results of the subjective assessment of task difficulty by the participants. Tasks were arranged on a scale from 1 (easiest) to 8 (most difficult).

Participants found basic website navigation and information retrieval to be the easiest tasks, except for finding contact information on Zara’s website. The most challenging tasks were those related to using the search engine, adding items to the cart, and managing the cart. This is also confirmed in the statistical analysis.

Table 5 focuses on the assessment of participants’ cognitive load while performing tasks, using a scale from 1 (very easy) to 10 (impossible to complete). The most difficult task involved setting filters, selecting item colors and sizes, and adding them to the cart. Similar to the first assessment, participants found tasks related to the search engine, cart management, and finding contact information on Zara’s website to be more cognitively demanding.

Table 4. Arrange the tasks from easiest to most difficult - participants’ answer (N = 4)

Part.	1		2		3		4		Mean $\pm$ SD		
Task	Zara	Zal.	Zara	Zal.	Zara	Zal.	Zara	Zal.	Zara	Zalando	Diff. (abs.)
1.	—	—	—	—	—	—	—	—	—	—	—
2.	1	1	2	2	1	1	1	2	1.25 ± 0.50	1.50 ± 0.58	0.25 ± 0.50
3.	2	2	1	1	2	2	2	1	1.75 ± 0.50	1.50 ± 0.58	0.25 ± 0.50
4.	8	8	8	7	7	8	8	7	7.75 ± 0.50	7.50 ± 0.58	0.25 ± 0.50
5.	7	7	7	8	5	6	5	8	6.00 ± 1.15	7.25 ± 0.96	1.25 ± 1.26
6.	4	5	4	5	4	5	4	3	4.00 ± 0.40	4.50 ± 1.00	0.50 ± 0.00
7.	6	6	6	6	8	7	6	6	6.50 ± 1.00	6.25 ± 0.50	0.25 ± 0.50
8.	3	4	3	4	3	4	3	5	3.00 ± 0.00	4.25 ± 0.50	1.25 ± 0.50
9.	5	3	5	3	6	3	3	4	4.75 ± 1.26	3.25 ± 0.50	1.50 ± 0.82

Table 5. Subjective assessment of participants’ mental workload (1 – very easy, 10 – unable to complete, N = 4)

Part.	1		2		3		4		Mean $\pm$ SD		
Task	Zara	Zal.	Zara	Zal.	Zara	Zal.	Zara	Zal.	Zara	Zalando	Diff. (abs.)
1.	—	—	—	—	—	—	—	—	—	—	—
2.	1	1	2	1	1	1	2	2	1.50 ± 0.58	1.25 ± 0.50	0.25 ± 0.50
3.	4	2	4	1	3	2	2	1	3.25 ± 0.96	1.50 ± 0.58	1.75 ± 0.96
4.	5	3	6	5	6	4	5	4	5.50 ± 0.58	4.00 ± 0.82	1.50 ± 0.58
5.	8	7	8	6	7	5	8	6	7.75 ± 0.50	6.25 ± 0.82	1.50 ± 0.50
6.	4	3	6	4	4	2	4	4	4.50 ± 1.00	3.25 ± 0.96	1.25 ± 0.96
7.	6	4	7	3	6	5	6	4	6.25 ± 0.50	4.00 ± 0.82	2.25 ± 1.26
8.	4	3	2	2	3	2	3	2	3.00 ± 0.82	2.25 ± 0.50	0.75 ± 0.50
9.	7	3	6	2	6	4	5	1	6.00 ± 0.82	2.50 ± 1.29	3.50 ± 1.00

Participants rated every task as causing higher cognitive load when using Zara’s website compared to Zalando’s website. This finding aligns fairly well with the time spent on each task by participants.

The statistical analysis using paired t-test was also performed for data obtained in the questionnaires related to subjective assessment of mental workload induced by each scenario task. The p-values are presented in Table 6. It can be observed that the significant statistical difference was found in case of tasks 3, 4, 5 and 7.

Table 6. The p-values calculated using paired t-test for each performed scenario task based on users' self-assessment scores ($N = 4$).

	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9
p-val	0.391	0.035	0.014	0.006	0.080	0.037	0.058	0.006

4 Discussion

The analysis of HRV parameters revealed notable differences between the resting period and tasks performed on the Zara and Zalando websites. Comparing the mean values, it was observed that mean heart rate was highest and HRV parameters were lowest when interacting with the Zara website. The lowest mean HR and highest HRV values were recorded during the resting period. These findings confirm the assumption that website accessibility influences CWL, with lower accessibility leading to higher CWL.

However, standard deviations were relatively high, particularly for the NN50 parameter, indicating substantial individual variation. When analyzing each participant separately, the observed trends between Zara and Zalando were consistent, but the same could not be said for the resting period, likely due to participants experiencing stress from the unfamiliar ECG measurement procedure.

Considering each task scenario separately, the mean values of HRV parameters further supported the assumption that CWL increases when working with a less accessible website. However, the high standard deviation values suggest that not all differences were statistically significant. The largest CWL differences were observed in tasks 3 and 7, aligning with self-assessed CWL ratings, where participants reported the most considerable differences.

A paired t-test analysis confirmed statistically significant differences in NNI, NN50, PNN50, and mean HR between tasks on Zara and Zalando websites. Additionally, significant differences were found for NNI, PNN50, and mean HR when comparing the resting period to Zara website tasks, but no significant differences were found between the resting period and Zalando tasks.

When analyzing individual tasks, the paired t-test results showed significant differences in NNI values for tasks 3, 5, and 6, SDNN values for tasks 3, 4, and 6, RMSSD values for tasks 5 and 7, NN50 values for task 4, PNN50 values for tasks 3, 4, and 7, and mean HR for tasks 3, 5, and 6. Additionally, statistical analysis of self-assessed CWL scores showed that the differences in mean scores

were significant for all tasks, except task 6, included in the HRV analysis, further supporting the link between website accessibility and cognitive workload.

The findings of this study align with previous research indicating that poor digital accessibility increases CWL and stress levels, as reflected in HRV measures. Studies have shown that decreased SDNN and RMSSD values are associated with heightened stress responses due to increased sympathetic nervous system activity, leading to lower cardiac variability [28,29]. The observed reduction in heart rate variability (HRV) while interacting with the poorly accessible Zara website aligns with existing research linking increased cognitive workload to diminished HRV. For instance, a study [30] demonstrated that HRV metrics effectively assess cognitive workload during task performance, highlighting HRV's sensitivity to mental demands. Additionally, a systematic review [31] emphasized HRV as a promising physiological correlate of cognitive functioning, reflecting autonomic control of the heart. These findings suggest that reduced digital accessibility can elevate cognitive demands, leading to measurable declines in HRV. The findings also confirm that higher cognitive demand is associated with lower HRV, as previously documented in workload studies [32]. Notably, high standard deviations, particularly in NN50, suggest considerable inter-individual differences, which may be linked to varying levels of adaptability and prior experience with assistive technologies [33]. These results emphasize the need for inclusive web design practices to mitigate cognitive strain for users with disabilities. Future studies should refine R-peak detection algorithms, integrate frequency-domain and nonlinear HRV parameters, and increase sample size to improve generalizability and reliability.

5 Conclusions

The aim of this study was to evaluate whether the level of website accessibility affects CWL of users with visual impairments, as determined by HRV analysis and NASA-TLX-inspired questionnaires. The research involved five visually impaired participants performing tasks on two websites with varying levels of accessibility: the Zara website, which demonstrated poor accessibility, and the Zalando website, which exhibited well-implemented accessibility features. The accessibility levels of these websites were assessed using the Markup Validation Service and the WAVE Web Accessibility Evaluation Tool and supported by expert evaluation. This pilot study suggests that poorly designed and inaccessible websites contribute to an increased cognitive workload, particularly for users with visual impairments. The HRV analysis demonstrated that interacting with a poorly accessible website leads to higher CWL, reflected in increased heart rate and decreased HRV parameters. The statistical analysis confirmed significant differences between tasks performed on the Zara and Zalando websites, reinforcing the impact of accessibility on cognitive demands. Future research will aim to improve the accuracy of R-peak detection algorithms, expand the analysis to include frequency-domain and nonlinear HRV parameters, and recruit a larger participant group to enhance the reliability and generalizability of the

findings. This research underscores the importance of prioritizing accessibility in web design to minimize cognitive strain for all users.

Acknowledgments. This work has been completed while the first author was the Doctoral Candidate in the Interdisciplinary Doctoral School at the Lodz University of Technology, Poland.

References

1. Sweller J (2011) Cognitive load theory. In: Alexander PA, Mayer R (eds) *Handbook of research on learning and instruction*. Routledge/Taylor & Francis Group, pp 37–76
2. Kosch T, Karolus J, Zagermann J, Reiterer H, Schmidt A, Woźniak PW (2023) A survey on measuring cognitive workload in human-computer interaction. *ACM Comput Surv* 55(13s):283. <https://doi.org/10.1145/3582272>
3. Sweller J (1988) Cognitive load during problem solving: effects on learning. *Cogn Sci* 12(2):257–285
4. World Health Organization: Disability. <https://www.who.int/news-room/fact-sheets/detail/disability-and-health>
5. World Health Organization: Blindness and vision impairment. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>
6. Das Chakladar D, Roy PP (2024) Cognitive workload estimation using physiological measures: a review. *Cogn Neurodyn* 18:1445–1465. <https://doi.org/10.1007/s11571-023-10051-3>
7. Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. *Adv Psychol* 52:139–183
8. Hedman L, Klingberg T, Enochsson L, Kjellin A, Felländer-Tsai L (2007) Visual working memory influences the performance in virtual image-guided surgical intervention. *Surg Endosc* 21:2044–2050
9. Kramer AF (2020) Physiological metrics of mental workload: a review of recent progress. *Multiple-Task Perform* 279–328 (2020)
10. Ma X, Monfared R, Grant R, Goh YM (2024) Determining cognitive workload using physiological measurements: pupillometry and heart-rate variability. *Sensors* 24(6):2010
11. Zhou Y, Huang S, Xu Z, Wang P, Wu X, Zhang D (2021) Cognitive workload recognition using EEG signals and machine learning: a review. *IEEE Trans Cogn Dev Syst* 14(3):799–818
12. Bear M, Connors B, Paradiso MA (2020) *Neuroscience: exploring the brain, Enhanced*. Jones & Bartlett Learning, Burlington
13. Bunce SC, Izzetoglu K, Ayaz H, Shewokis P, Izzetoglu M, Pourrezaei K (2011) Implementation of fNIRS for monitoring levels of expertise and mental workload. In: *International conference on foundations of augmented cognition*. Springer, New York, pp 13–22
14. Hirachan N, Mathews A, Romero J, Rojas RF (2022) Measuring cognitive workload using multimodal sensors. In: *2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC)*. IEEE, pp 4921–4924
15. Yang S et al (2021) The impacts of temporal variation and individual differences in driver cognitive workload on ECG-based detection. *Hum Factors* 63(5):772–787

16. Chikhi S, Matton N, Blanchet S (2022) EEG power spectral measures of cognitive workload: a meta-analysis. *Psychophysiology* 59(6):e14009
17. Kalatzis A, Teotia A, Prabhu VG, Stanley L (2021) A database for cognitive workload classification using electrocardiogram and respiration signal. In: Ayaz H, Asgher U, Paletta L (eds) *AHFE 2021*, vol 259. LNNS. Springer, Cham, pp 509–516. https://doi.org/10.1007/978-3-030-80285-1_58
18. Albers MJ (2011) Tapping as a measure of cognitive load and website usability. In: *Proceedings of the 29th ACM international conference on Design of communication*, pp 25–32
19. Tracy JP, Albers MJ (2006) Measuring cognitive load to test the usability of web sites. In: *Annual conference-society for technical communication*, vol 53, p 256
20. Hewitt DH, He Y (2022) Cognitive load and website usability: effects of contrast and task difficulty. In: *Proceedings of the human factors and ergonomics society annual meeting*, vol 66, no 1. SAGE Publications, Los Angeles, pp 1809–1813
21. Abdelaal Y, Al-Thani D (2023) Accessibility first: detecting frustration in web browsing for visually impaired and sighted smartphone users. *Univ Access Inf Soc*. <https://doi.org/10.1007/s10209-023-01053-3>
22. Caldiroli CL et al (2023) Comparing online cognitive load on mobile versus PC-based devices. *Pers Ubiquit Comput* 27(2):495–505
23. Solhjoo S, Haigney MC, McBee E et al (2019) Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Sci Rep* 9:14668. <https://doi.org/10.1038/s41598-019-50280-3>
24. Muthukrishnan SP, Gurja JP, Sharma R (2017) Does heart rate variability predict human cognitive performance at higher memory loads. *Indian J Physiol Pharmacol* 61(1):14–22
25. Ling Z, Xu-Qiang Q, Tao NI, Wei Y, Yi-Nong L (2021) Driver cognitive loads based on multi-dimensional information feature analysis. *China J Highway Transp* 34(4)
26. World Wide Web Consortium: Web Content Accessibility Guidelines (WCAG) 2.1. <https://www.w3.org/TR/WCAG21/>
27. World Wide Web Consortium: Website Accessibility Conformance Evaluation Methodology (WCAG-EM) 1.0. <https://www.w3.org/TR/WCAG-EM/>
28. Laborde S, Mosley E, Thayer JF (2017) Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. *Front Psychol* 8:213. <https://doi.org/10.3389/fpsyg.2017.00213>
29. Shaffer F, Ginsberg JP (2017) An overview of heart rate variability metrics and norms. *Front Public Health* 5:258. <https://doi.org/10.3389/fpubh.2017.00258>
30. Digiesi S, Manghisi VM, Facchini F, Klose EM, Foglia MM, Mummolo C (2020) Heart rate variability based assessment of cognitive workload in smart operators. *Manag Prod Eng Rev* 11(3):56–64
31. Forte G, Favieri F, Casagrande M (2019) Heart rate variability and cognitive function: a systematic review. *Front Neurosci* 13:710
32. Cinaz B, Arnrich B, La Marca R, Tröster G (2013) Monitoring of mental workload levels in an office environment using physiological sensors. In: *Proceedings of the ACM conference on pervasive and ubiquitous computing*, pp 145–154. <https://doi.org/10.1145/2493432.2493441>
33. Fairclough SH, Gilleade K (eds) (2014) *Advances in Physiological Computing*. HIS. Springer, London. <https://doi.org/10.1007/978-1-4471-6392-3>



Evaluation of Factors-of-Interest in Muscle Phantoms Based on Ultrasonic Data

Vladislavs Agarkovs¹, Aleksandrs Sisojevs^{1,2} , Alexey Tatarinov¹ ,
and Tamara Laimiņa¹

¹ Institute of Electronics and Computer Science, 14 Dzerbenes Street, Riga 1006, Latvia
alexiv@inbox.lv

² Riga Technical University, 6A Kipsalas Street, Riga 1048, Latvia

Abstract. Subcutaneous adipose tissue (SAT) and intermuscular adipose tissue (IMAT) are two forms of body fat. The differential assessment of SAT and IMAT contents in sarcopenia and obesity by ultrasonic measurements is difficult due to their complex influence. In the present study, the possibility of using pattern recognition methods applied to ultrasonic measurement data to discriminate SAT and IMAT values was investigated on muscle tissue mimicking phantoms.

The muscle mimicking phantoms were made of gelatin (muscle) with oil (fat) layers (SAT) and inclusions (IMAT) with gradually varying SAT and IMAT contents from zero to 50% with a step 12.5%. The set included a net of 25 phantoms with five grades of SAT and IMAT, each varied independently.

Each phantom (object under study) passed multiple repeatable ultrasound tests. For each object, a set of parameters was extracted from the set of ultrasonic signals, including speed of sound, ultrasound attenuation at two frequencies and signal amplitude integrals. Decision rules were constructed for all objects based on the set of the extracted parameters. The efficiency of the constructed rules was checked using the signals of the examination object. The results of the experiments demonstrated the potential effectiveness of the proposed approach for determining SAT and IMAT values separately.

Keywords: Quantitative Ultrasound · Pattern Recognition · Tissue Models

1 Introduction

Subcutaneous adipose tissue (SAT) and intermuscular adipose tissue (IMAT) are two forms of body fat. SAT is the most widely distributed subcutaneous fat tissue layer composed of adipocytes is found just beneath the skin. IMAT is typically the broadest definition of fatty infiltration in the muscle referring to storage of lipids in adipocytes underneath the deep fascia of muscle. This includes the visible storage of lipids in adipocytes located between the muscle fibers and between muscle groups [1].

Independent evaluation of SAT and IMAT contents has become crucial as independent health risk factors during last decades is by several reasons. In western countries, the increasing proportion of people over age 60 is linked to a high rate of obesity that associates with frailty syndrome [2]. Muscle mass decreases through the years, whereas

SAT mass increases until midlife. Elderly age is associated with decrease in SAT (parallel to the loss of lean body mass) and the emergence of ectopic fat in muscle. In the core of both frailty and sarcopenia there are dysfunction and deterioration of the muscle and the fat tissues and fat tissue accumulation. Muscle mass loss and fat accumulation in the muscle in the elderly, with or without the presence of obesity, explain the frail in sarcopenic population. Osteosarcopenia linking atrophies of bone and muscle is a newly defined geriatric syndrome with serious consequences in the older population. Studies of pathogenesis of osteosarcopenia in the light of new discoveries underline the importance of local lipotoxicity induced by IMAT [3]. Other studies revealed independent association of IMAT deposit in the thigh with cardiometabolic risk factors (including total blood cholesterol, low-density lipoprotein (LDL), and triglycerides), and decreased insulin sensitivity. The SAT/IMAT ratio is associated with serum adiponectin levels and this factor could possibly serve as an indicator for individual factors of a person's cardiovascular risk [4]. Among systemic adipose tissues, SAT significantly and negatively associate with muscle mass and forces, and could be related to the presence and progression of knee osteoarthritis [5].

Body composition assessments vary in precision and in the target tissue of interest, including anthropometric measurements, bioelectrical impedance, dual-energy X-ray absorptiometry and body density [6]. There is no single universally recommended method for body composition assessment in the obese, but each modality has benefits and drawbacks. Computed tomography (CT) and magnetic resonance imaging (MRI) are gaining in popularity for body composition assessment, especially effective in SAT and IMAT segmentation. MRI is the dominant 3D imaging modality to quantify muscle properties in skeletal muscle disorders, in inherited and acquired muscle diseases, sarcopenia, and frailty [7]. In combination with appropriate segmentation, T1w images are used for the quantitative measurement of area or volume of tissue compartments such as SAT, IMAT and muscle volume. The advantage of CT is the possibility to quantify the muscle density also termed as muscle attenuation or muscle radiation attenuation, which linearly depends on the muscle fat content [8]. However, stationary and costly equipment, its inaccessibility, as well as the requirement for special qualifications of personnel make mass use of MRI and CT difficult. Ultrasonography or clinical ultrasound, exploring the associations between skeletal muscle echogenicity and its physical properties presents an opportunity to be utilized as a screening tool [9]. It suggests quantification of SAT thickness; however, the prediction of IMAT is still problematic.

Ultrasonometry, which offers an assessment of the SAT and IMAT content based on known dependencies of the acoustic parameters of the tissue (velocity and frequency function of ultrasound attenuation) on the fat content [10], may be the most convenient and efficient modality for diagnostics. However, at the present, the possibility of separating SAT and IMAT based on ultrasound measurements seems problematic. The aim of the study was to demonstrate the potential for differential assessment of SAT and IMAT in simple muscle models by applying pattern recognition methods to ultrasound signals.

Similar approaches of pattern recognition applications for the evaluation of factors-of-interest based on ultrasonic parameters as decision rules have been developed previously for different physical objects and corresponding testing tasks. For example, the possibility of determining the depth and the degree of deterioration of the surface layer

of concrete by the combined use of ultrasonic surface waves at different frequencies was proven implementing a mathematical approach of pattern recognition with statistical criteria [11]. The universality of this approach for testing different materials and assessment of different factors-of-interest was shown [12]. The estimation method for cortical thickness and intracortical porosity in compact bone models of osteoporosis was based on a statistical analysis of the magnitude of the DFT spectrum of ultrasonic signals and physical parameters of ultrasound propagation [13–15].

2 Methods

2.1 Muscle Phantoms with SAT and IMAT Variation

The muscle mimicking phantoms were made of gelatin (muscle) with oil (fat) layers (SAT) and inclusions (IMAT) with gradually varying SAT and IMAT contents from zero to 50% with a step 12.5%. The set included a net of 25 phantoms with five grades (0–4) of SAT and IMAT that varied independently presenting 25 possible combinations of SAT and IMAT (Fig. 1). The phantoms presented rectangular boxes with a fixed acoustical distance 80 mm and acoustically transparent windows for transducers. Parallel SAT layers on the both sides of the phantoms were orthogonal to the acoustic path, while IMAT inclusions were 5 mm holes in a checkerboard pattern in the volume of gelatin.

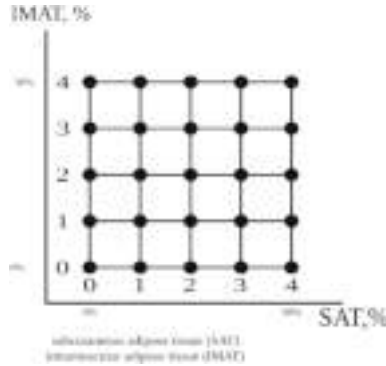


Fig. 1. Set of phantoms with gradually varied SAT and IMAT contents.

2.2 Parameter Extraction

Trains of ultrasonic signals were acquired by through transmission using a pair of emitting and receiving transducers excited by tone-bursts. The signals were digitized at sampling rate 30 MHz with a 10-bit ADC. Each signal train consisted of three signals at frequencies configurations 1 MHz, 2 MHz and sweep between 0.5 and 2.5 MHz (Fig. 2). Each phantom was measured 15 times that resulted in 375 individual signal sets.



Fig. 2. Example of train of received ultrasonic signals in a single run.

From each signal train, six parameters were extracted:

1) Speed of sound. Speed of sound S (m/s) was calculated as

$$S = \frac{L \cdot f}{t^* - \Delta t}, \quad (1)$$

where L is distance between emitter and receiver, f is sampling rate, t^* is ultrasound arrival time in data points, Δt is transducers constant at zero distance;

- 1) Ultrasound attenuation at 1 MHz frequency;
- 2) Ultrasound attenuation at 2 MHz frequency;
- 3) Ultrasound attenuation at a certain frequency α (dB/cm) was calculated as

$$\alpha = \frac{(0.021 \log(I_0/I_s))}{L}, \quad (2)$$

where I_0 is integral of the signal amplitudes at zero distance, I_s is integral of the signal amplitudes at distance L .

- 4) Attenuation ratio at 1 and 2 MHz, determined as $\alpha_{1\text{MHz}} / \alpha_{2\text{MHz}}$;
- 5) Integral of sweep signal;
- 6) Ratio of integrals of direct sweep signal and its third reflection (Fig. 2)

Integrals were calculated as sum of all amplitude values in a selected signal interval:

$$I = \sum_{n_s}^{n_e} |S_n|, \quad (3)$$

where n_s is start index, n_e is end index, $S_n - n_{th}$ is data point in signal.

2.3 Building of Decision Rules

After calculating these parameters for each signal in each phantom, the maximum and minimum values for each parameter for each phantom were calculated. The maximum - minimum values were used to construct a three-dimensional volume, where the upper boundary surface was a piecewise linear interpolation function with the maximum parameter value and the lower boundary surface was a piecewise linear interpolation function with the minimum parameter value.

Then, to increase the number of data points in this volume, both surfaces, the top and bottom, were interpolated using Akima interpolation, and 3 new interpolated points

were added between each real data point. This increased the number of points from 5 per axis to 17. The x-axis is the SAT value for each sample, and the y-axis is the IMAT value for each sample. Example of volume for the first parameter, speed of sound, is shown in Fig. 3.

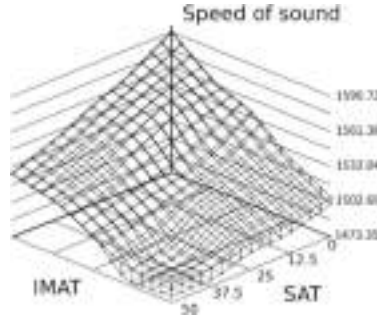


Fig. 3. Example of constructed volume for speed of sound as a complex function of SAT and IMAT.

2.4 Evaluation of Unknown Object

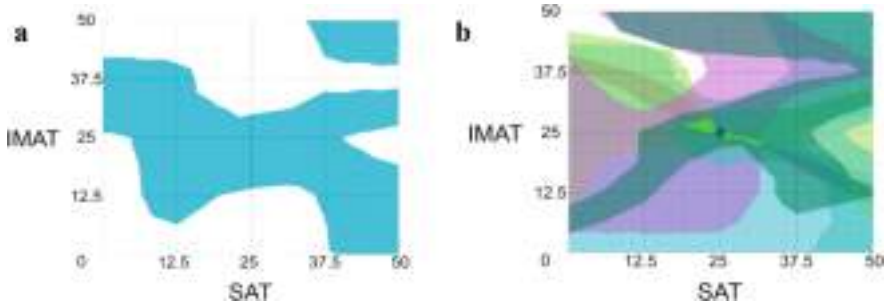


Fig. 4. Example of intersection of unknown parameter value with constructed parameter volume (a) and example of superposition of multiple parameters intersection areas (b). Different areas are denoted with different colors; projected answer is indicated as blue circle, evaluated answer is indicated as green circle.

To evaluate the position of a new object (phantom) with unknown SAT and IMAT values in the SAT-IMAT coordinate field, firstly, the same parameters are to be extracted from the train of ultrasonic signals. Then, an intersection area can be created between the extracted parameter and the volume of the same parameter, resulting in one or more intersection areas (Fig. 4a). This intersection area represents a set of possible answers. To get the final answer, it is necessary to find the intersection between all six areas corresponding to six decision rules (circled in green in Fig. 4b).

3 Results of Evaluation

To evaluate this method, the input data were split into training data and test data. From each measured object containing 15 signal trains obtained by repeated measurements, 10 signals trains were used to construct the decision rule, and 5 signal trains were used for estimation. The signal trains in each object were selected randomly.

The results are presented in Fig. 5 as differences between the correct (projected) and estimated values for SAT and IMAT, sorted in ascending order of error (difference). SAT and IMAT contents are denoted in the names of objects in Fig. 5. For example, 12.5SAT_50IMAT means that the object contains 12.5% SAT and 50% IMAT.

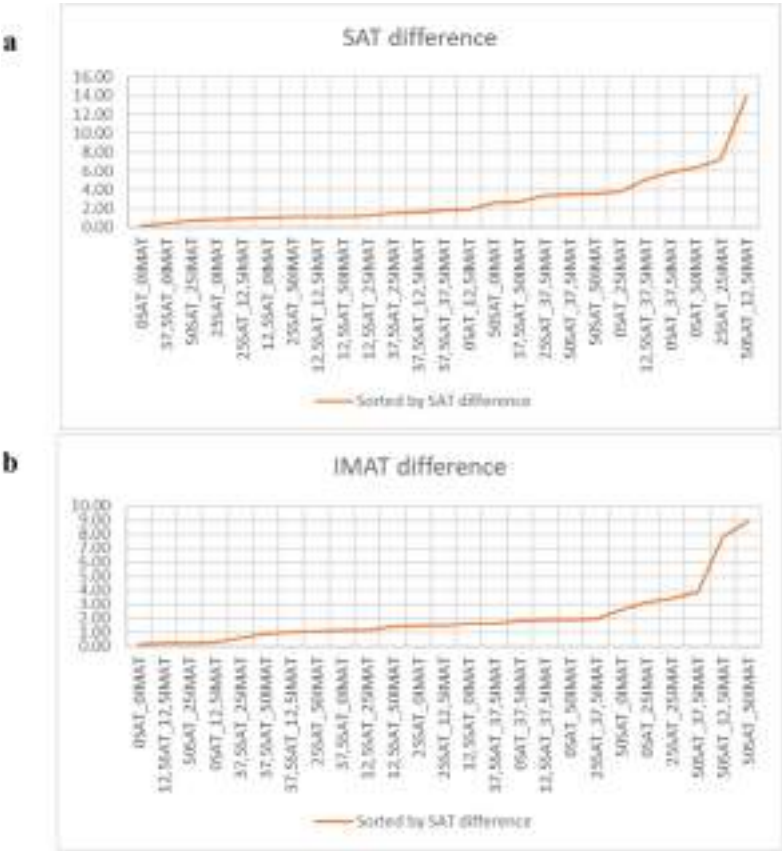


Fig. 5. Evaluation results for SAT (a) and IMAT (b).

These results showed that IMAT evaluation error is below 5% for 94% of the objects and SAT evaluation error is below 5% for 80% of the objects. Average errors are 2.94% for SAT and 2.07% for IMAT. Errors exceeding 5% are typical for objects with high SAT and IMAT contents, i.e., if we transfer these results to human subjects, these are individuals prone to obesity.

4 Conclusions

1. The results demonstrated the fundamental feasibility of the proposed approach based on ultrasonic signals and pattern recognition to distinguishing between SAT and IMAT contents in human extremities (lower leg, thigh) with diagnostically acceptable accuracy.
2. In get closer to diagnostic conditions as the next steps, it is necessary to transfer the method to data obtained on human subjects, where SAT and IMAT would be assessed by acknowledged imaging modalities (MRI, CT).
3. To increase the reliability of diagnostics, it is worth to both develop additional informative decision rules and prioritize them.

Acknowledgement. The study was supported by the research project of the Latvian Council of Science lzp-2021/1–0290 “Comprehensive assessment of the condition of bone and muscle tissues using quantitative ultrasound” (BoMUS).

References

1. Addison O, Marcus RL, Lastayo PC, Ryan AS (2014) Intermuscular fat: a review of the consequences and causes. *Int J Endocrinol* 2014:309570. <https://doi.org/10.1155/2014/309570>
2. Buch A et al (2016) Muscle function and fat content in relation to sarcopenia, obesity and frailty of old age—an overview. *Exp Gerontol* 76:25–32. <https://doi.org/10.1016/j.exger.2016.01.008>
3. Ghasemikaram M et al (2021) Effects of 16 months of high intensity resistance training on thigh muscle fat infiltration in elderly men with osteosarcopenia. *Geroscience* 43(2):607–617. <https://doi.org/10.1007/s11357-020-00316-8>
4. Hassler EM et al (2021) Distribution of subcutaneous and intermuscular fatty tissue of the mid-thigh measured by MRI-A putative indicator of serum adiponectin level and individual factors of cardio-metabolic risk. *PLoS ONE* 16(11):e0259952. <https://doi.org/10.1371/journal.pone.0259952>
5. Chang J, Liao Z, Lu M, Meng T, Han W, Ding C (2018) Systemic and local adipose tissue in knee osteoarthritis. *Osteoarthritis Cartilage* 26(7):864–871. <https://doi.org/10.1016/j.joca.2018.03.004>
6. Duren DL et al (2008) Body composition methods: comparisons and interpretation. *J Diabetes Sci Technol* 2(6):1139–1146. <https://doi.org/10.1177/193229680800200623>
7. Engelke K et al (2023) Magnetic resonance imaging techniques for the quantitative analysis of skeletal muscle: state of the art. *J Orthop Translat* 19(42):57–72. <https://doi.org/10.1016/j.jot.2023.07.005>
8. Engelke K, Museyko O, Wang L, Laredo JD (2018) Quantitative analysis of skeletal muscle by computed tomography imaging-state of the art. *J Orthop Translat* 28(15):91–103. <https://doi.org/10.1016/j.jot.2018.10.004>
9. Oranchuk DJ, Bodkin SG, Boncella KL, Harris-Love MO (2024) Exploring the associations between skeletal muscle echogenicity and physical function in aging adults: a systematic review with meta-analyses. *J Sport Health Sci* 13(6):820–840. <https://doi.org/10.1016/j.jshs.2024.05.005>

10. Koch T, Lakshmanan S, Brand S, Wicke M, Raum K, Mörlein D (2011) Ultrasound velocity and attenuation of porcine soft tissues with respect to structure and composition: I. Muscle. *Meat Sci* 8(1):51–58. <https://doi.org/10.1016/j.meatsci.2010.12.002>
11. Tatarinovs A, Sisojevs A, Chaplinska A, Shahmenko G, Kurtenoks V (2022) An approach for assessment of concrete deterioration by surface waves. *Procedia Struct Integrity* 37:453–461. ISSN 2452-3216. <https://doi.org/10.1016/j.prostr.2022.01.109>
12. Sisojevs A, Tatarinov A, Kovalovs M, Krutikova O, Chaplinska A (2022) An approach for parameters evaluation in layered structural materials based on DFT analysis of ultrasonic signals. In: *Proceedings of the 11th international conference on pattern recognition applications and methods ICPRAM'2022*. SciTePress, pp 307–314. ISBN 978-989-758-549-4; ISSN 2184-4313. <https://doi.org/10.5220/0010878400003122>
13. Sisojevs A, Tatarinov A, Chaplinska A (2023) Evaluation of factors-of-interest in bone mimicking models based on DFT analysis of ultrasonic signals. In: *Proceedings of the 12th international conference on pattern recognition applications and methods ICPRAM'2023*. SciTePress, pp 914–919. ISBN 978–989–758–626–2; ISSN 2184-4313. <https://doi.org/10.5220/0011742800003411>
14. Tatarinov A, Sisojevs A, Chaplinska A (2023) Pattern recognition based approach for extraction of factors of interest from ultrasonic data. In: *International workshop on embedded digital intelligence (IWEDI'2023)*. EDI
15. Mamontovs M, Sisojevs A, Tatarinov A (2024) Bone porosity and thickness evaluation by ultrasonic signals using pattern recognition. In: *Proceedings of the international conferences on interfaces and human computer interaction 2024, game and entertainment technologies 2024 and computer graphics, visualization, computer vision and image processing 2024 – part of the 18th multi conference on computer science and information systems 2024*, Budapest, Hungary, 13–15 July 2024. IADIS Press, pp 200–204. ISBN: 978-989-8704-58-0



Nonlinear Analysis of ECG Signals in Arrhythmia Simulated by MPS450 Fluke and Measured with ADS1298 AFE

Elzbieta Raus-Jarzabek¹ and Elzbieta Olejarczyk^{1,2}(✉)

¹ AGH University of Krakow, Al. Mickiewicza 30, 30-059 Krakow, Poland
eolejarczyk@agh.edu.pl

² Nalecz Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences,
Trojdena 4 Str., 02-109 Warsaw, Poland

Abstract. Arrhythmia detection and classification are critical for early diagnosis and management of cardiac disorders. In this study, we performed a nonlinear analysis of ECG signals representing up to 35 types of arrhythmias generated by the MPS450 Fluke simulator. These signals were recorded with the ADS1298 analog front-end (AFE). Eight measures were analysed such as variance, Higuchi Fractal Dimension (*HFD*), Katz Fractal Dimension (*KFD*), Detrended Fluctuation Analysis (*DFA*), Sample Entropy (*SampEn*), Approximate Entropy (*ApEn*), and Multiscale Entropy (*MSE*). The impact of window size was tested also. These measures allowed to differentiate between a normal rhythm and 20 out of 35 arrhythmias as well as between 73.7% of pairs of different arrhythmias, highlighting their diagnostic potential. The best results were obtained for 30-s window. All fractal dimension measures (*KFD*, *HFD* and *DFA*) and *ApEn* were the most important features for arrhythmias classification. This study demonstrates the efficacy of nonlinear ECG signal analysis in characterizing arrhythmias and emphasizes the importance of advanced simulation and acquisition systems like the MPS450 Fluke and ADS1298 for cardiac research and clinical applications, i.e. integration of an application based on nonlinear measures with wearable ECG devices for automated arrhythmia classification.

Keywords: ECG · arrhythmia · nonlinear analysis · MPS450 Fluke simulator · ADS1298

1 Introduction

Cardiovascular diseases are the leading cause of death worldwide, and arrhythmias are a major contributor [1]. Arrhythmias referring to irregularity of the heartbeat can range from harmless to life-threatening. While some arrhythmias may cause no symptoms, others can lead to palpitations, dizziness, fatigue, chest pain, or even cardiac arrest [2]. The cost of treatment, including diagnostics, medication, and interventions like catheter ablation, places a significant strain on healthcare systems. Improving arrhythmia detection and management has the potential to save millions of lives annually and to reduce the global healthcare burden.

The fractal dimension and entropy measures were widely applied in the nonlinear analysis of heart rate variability (HRV). The fractal measures such as Higuchi fractal dimension (*HFD*) and Katz fractal dimension (*KFD*), quantifying the signal complexity, are particularly effective for detecting arrhythmias and assessing HRV under different physiological and pathological conditions [3]. Detrended Fluctuation Analysis (*DFA*) has emerged as a robust tool for analysing long-term correlations in nonstationary ECG signals [4]. The scaling exponent ranges from 0.9 to 1.19 for healthy individuals, while those with arrhythmias such as premature ventricular contractions (PVCs) often show significantly lower values, typically around 0.7 or even lower in severe cases [5]. A modified *DFA* allows for detection of severe cardiac abnormalities with scaling indices approaching 0.5 in extreme conditions [6]. Entropy-based methods quantify the irregularity and unpredictability of ECG signals. Sample Entropy (*SampEn*) is consistently higher in individuals with atrial fibrillation (AF) compared to those without arrhythmias. For example, in a study on 15 chronic AF patients, *SampEn* values of 0.392 for time lag equal to 1 and 0.759 for time lag equal to 5 were significantly higher than the values of 0.252 and 0.533 observed in subjects without AF [7]. Additionally, *SampEn* can be applied to predict AF recurrence following ablation, with elevated values of 1.11 in patients experiencing recurrence in relation to 0.91 in those maintaining sinus rhythm [8]. Another study highlighted the utility of Approximate Entropy (*ApEn*) in atrial fibrillation, reporting a mean *ApEn* of 0.556 ± 0.169 in chronic AF patients which is significantly higher than the 0.347 ± 0.119 observed in healthy subjects [9]. Similarly, untreated AF patients exhibited higher *ApEn* values (0.687 ± 0.033) compared to those on anti-arrhythmic drugs (0.516 ± 0.209). An evaluation of entropy across multiple time scales by Multiscale Entropy (*MSE*) offers a possibility to study both short-term and long-term dynamics [10]. Some researchers indicated the usefulness of all entropy measures for classification of arrhythmias such as atrial fibrillation and ventricular tachycardia in clinical diagnostics [11–14].

The integration of fractal and entropy-based features offers a robust framework for ECG analysis. A short computational time of these measures makes them suitable to apply in wearable devices for real-time arrhythmia monitoring [15]. While each metric has its limitations, their combination can enhance diagnostic accuracy.

The choice of an appropriate Analog Front-End (AFE) can enhance additionally the detection performance. The ADS1298 Analog Front-End (AFE) produced by Texas Instruments Inc. is designed for high-resolution ECG acquisition. The ADS1298 has many features useful in acquisition of ECG signals such as: 24-bit resolution for capturing subtle signal variations and high sampling frequency (1 kHz) [16].

The use of application based on nonlinear ECG analysis integrated with technology of ADS 1298 may be a powerful diagnostic tool for detection of arrhythmias.

In this study we propose to apply a set of eight nonlinear features directly to ECG signals and not to HRV as it has been used by most researchers so far. We hypothesises that such approach will allow for differentiation of a wider range of arrhythmias and, what is important in ECG monitoring, for performing the calculations in a shorter time window. A comprehensive analysis of the full range of different arrhythmias was possible using the MPS450 Fluke simulator.

2 Materials and Methods

2.1 Data Simulation

The ECG signals used in this study were simulated using the MPS450 Fluke Biomedical Patient Simulator, a high-precision device allowing for generation of various physiological and pathological cardiac rhythms, including 35 distinct arrhythmia types [17]. The following arrhythmias were considered: (1) coarse atrial fibrillation; (2) fine atrial fibrillation; (3) atrial flutter; (4) sinus arrhythmia; (5) missed beat; (6) atrial tachycardia; (7) paroxysmal atrial tachycardia; (8) nodal rhythm; (9) supraventricular tachycardia; (10) premature atrial contraction (PAC); (11) nodal PAC; (12) premature ventricular contraction 1 (PVC1) left ventricular (LV); (13) PVC1 LV early; (14) PVC1 LV R on T; (15) PVC2 right ventricular (RV); (16) PVC2 RV early; (17) PVC2 RV R on T; (18) PVCs 6/MIN; (19) PVCs 12/MIN; (20) PVCs 24/MIN; (21) frequent multifocal PVCs; (22) bigeminy; (23) trigeminy; (24) pair PVCs; (25) run of 5 PVCs; (26) run of 11 PVCs; (27) ventricular tachycardia; (28) ventricular fibrillation coarse; (29) ventricular fibrillation fine; (30) asystole; (31) 1st degree block; (32) 3rd degree block; (33) right bundle branch block (RBBB) block; (34) left bundle branch block (LBBB); (35) 2nd degree block.

An example of simulated ECG signals are shown in Fig. 1 for regular rhythm and for four types of common arrhythmias: atrial fibrillation, ventricular tachycardia, ventricular fibrillation, and premature ventricular contractions. The simulated signals closely mimic real-world conditions, providing a reliable dataset for nonlinear analysis.

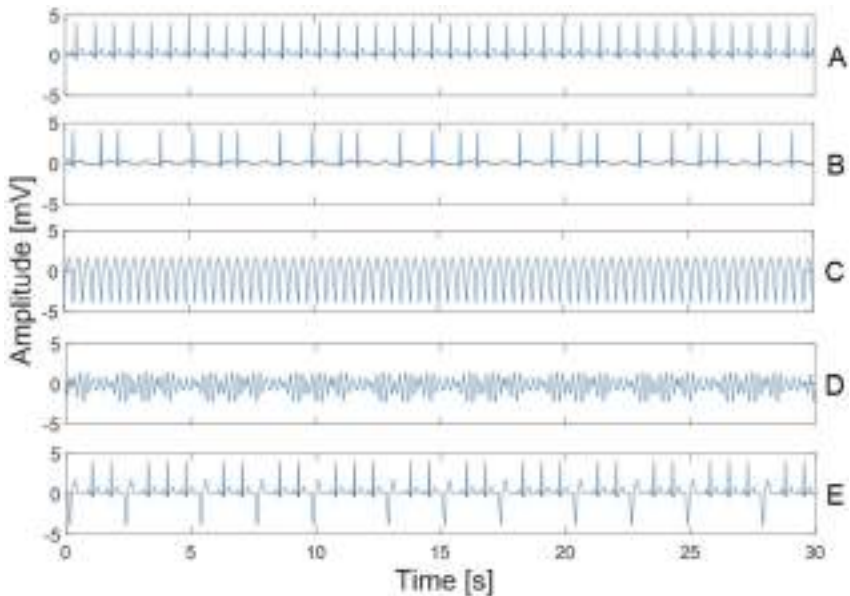


Fig. 1. An example of 30-s ECG for a regular rhythm (A) and for four types of arrhythmia: atrial fibrillation (B), ventricular tachycardia (C), ventricular fibrillation (D), and premature ventricular contractions (E), simulated by MPS450 Fluke simulator and recorded by ADS1298.

These signals were recorded using the ADS1298 Analog Front-End (AFE). The 5-min ECG signals were acquired on lead II, commonly used for arrhythmia analysis. The sampling frequency was 1000 Hz. The signals were filtered using a 50 Hz notch filter and bandpass Butterworth filter in frequency range from 0.5 to 70 Hz. Then, the signal was normalized based on maximal and minimal value difference.

2.2 Nonlinear Analysis

The ECG analysis was performed used eight measures: variance, Higuchi's Fractal Dimension (*HFD*) [18], Katz's Fractal Dimension (*KFD*) [19], Detrended Fluctuation Analysis (*DFA*) [20], Sampling Entropy (*SampEn*) [21], Approximate Entropy (*ApEn*) [22], and Multiscale Entropy (*MSE*) calculated for three scales (*MSE1*, *MSE2*, *MSE3*) [23].

Each measure has its own specific characteristics. Below we are providing a short description of these measures.

Variance quantifies the dispersion of data points in a signal around its mean.

$$\text{Variance} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (1)$$

where x_i are signal values, μ is the mean of the signal, and N is the number of samples.

HFD estimates the fractal dimension of a time series by analyzing its self-similarity over different scales. This is a measure of the ECG signals complexity. The only parameter is k . In our study k is equal to 8.

The algorithm consists in the following steps:

1. Divide the time series into segments of length varying from 1 to k .
2. Calculate the average curve length L_k for each segment size according the formula:

$$L_k = \frac{1}{k} \sum_{m=1}^k \left(\frac{\sum_{i=1}^{n_m} |x(m + i \cdot k) - x(m + (i - 1) \cdot k)|}{n_m \cdot k} \right) \quad (2)$$

where n_m is the number of data points in the m -th segment.

3. Perform a log-log regression of L_k against $1/k$. The slope of the regression line gives the fractal dimension.

KFD calculates the geometric irregularity of a signal based on its total curve length and the planar extent.

$$\text{KFD} = \frac{\log_{10}(n)}{\log_{10}(n) + \log_{10}(d/L)} \quad (3)$$

where n is the number of points, d is the maximum distance from the starting point to any other point of the sequence, and L is a sum of Euclidean distances between successive points.

DFA evaluates long-range correlations in a time series by analyzing the fluctuations of detrended data over different segment lengths.

The algorithm consists in the following steps:

1. Divide the signal x into non-overlapping segments of length N for each of n scales.
2. Detrend each segment using linear regression calculated as a polynomial of order equal to 1.
3. Compute the root-mean-square fluctuation for each segment:

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^N [x(i) - y(i)]^2} \quad (4)$$

where $y(i)$ is the fitted trend.

4. Plot $\log(F(n))$ versus $\log(n)$. The slope of the line is the scaling exponent.

SampEn measures the regularity and predictability of a signal x by quantifying the likelihood that sequences of length m (here m is set to 2) match within a tolerance r to $m + 1$ points.

$$\text{SampEn}(m, r, N) = -\ln\left(\frac{A}{B}\right) \quad (5)$$

where A is the number of matched sequences of length $m + 1$; B is the number of matched sequences of length m ; $r = 0.2 \cdot \text{std}(x)$.

ApEn quantifies the unpredictability of fluctuations in a signal by evaluation of the logarithmic probability of similar patterns occurring.

$$\text{ApEn}(m, r, N) = \varnothing^m(r) - \varnothing^{m+1}(r) \quad (6)$$

where:

$$\varnothing^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln(C_i^m(r)) \quad (7)$$

and $C_i^m(r)$ is the number of patterns similar to a given pattern within a tolerance r .

MSE extends *SampEn* by evaluating signal complexity across multiple scales. The algorithm consists in the following steps:

1. Coarse-grain the signal at each scale by averaging consecutive non-overlapping segments.
2. Compute *SampEn* for each coarse-grained signal.
3. Analyze the entropy values across scales.

The entropy measures (*SampEn*, *ApEn* and *MSE*) were computed using the Entropy-Hub toolbox in Matlab, while fractal dimension measures (*HFD*, *KFD* and *DFA*) using a custom Matlab functions.

2.3 Statistical Analysis

The Kruskal-Wallis test with multiple comparisons at $p\text{-value} < 0.05$ for individual measures calculated in a 30-s window was applied to compare 35 arrhythmias with each other and with normal rhythm.

3 Results

3.1 Dependence of Nonlinear Measures on Window Size

All measures were calculated for four window sizes: 5, 10, 20 and 30 s. The feature values depend on window size. Generally, the smallest standard deviations were observed for longer windows (c.f. Fig. 2). Thus, the 30-s window was chosen for further analysis.

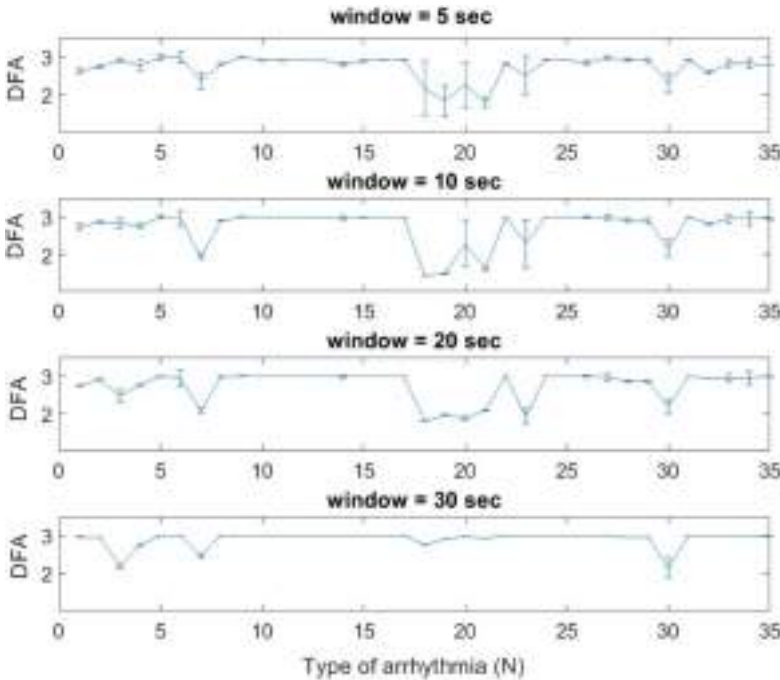


Fig. 2. Mean *DFA* of ECG in different types of arrhythmia calculated in function of window size. The standard deviation of *DFA* is marked with the vertical bars.

3.2 Differentiation Between Regular Rhythm and Arrhythmia

The significant differences were found between regular rhythm and 20 out of 35 types of arrhythmias (see Table 1). The lack of statistically significant differences between the values of the measures for regular rhythm and arrhythmia are found for $N = 5, 10\text{--}17, 24\text{--}26, 31$ and 35 (marked in orange in Table 1). This group comprises: missed beat, all types of premature atrial and ventricular contractions, and 1st and 2nd degree block.

3.3 Differentiation Between Different Types of Arrhythmia

The *p-value* matrices are presented separately for each measure in Fig. 3. Zero values were assigned to *p-values* below 0.05 (marked in black in Fig. 3) and one to *p-values*

Table 1. Mean values of nine measures (*var*, *HFD*, *KFD*, *DFA*, *SampEn*, *ApEn*, *MSE1*, *MSE2*, *MSE3*) for 35 types of arrhythmia and for regular rhythm ($N = 36$) ordered according to the importance of measure in differentiating arrhythmia from regular rhythm. The significant differences between regular rhythm and an appropriate arrhythmia were marked in red.

N	<i>std</i>	<i>HFD</i>	<i>KFD-1</i>	<i>DFA</i>	<i>SampEn</i>	<i>ApEn</i>	<i>MSE1</i>	<i>MSE2</i>	<i>MSE3</i>
36	3,41E-07	1,05	1,69E-09	3,00	0,96	1,44	2,96	2,64	2,59
5	3,39E-07	1,08	6,50E-09	3,00	0,96	1,44	2,46	2,73	2,63
10	3,41E-07	1,04	1,69E-09	3,00	0,96	1,45	2,99	2,57	2,62
11	3,41E-07	1,04	1,69E-09	3,00	0,96	1,45	2,99	2,57	2,62
12	3,41E-07	1,04	1,69E-09	3,00	0,96	1,45	3,00	2,57	2,58
13	3,41E-07	1,04	1,69E-09	3,00	0,96	1,45	2,99	2,57	2,56
14	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	3,01	2,58	2,55
15	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	2,86	2,44	2,52
16	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	2,87	2,43	2,48
17	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	2,99	2,58	2,63
24	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	3,06	2,55	2,60
25	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	3,00	2,45	2,44
26	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	2,99	2,44	2,51
31	3,41E-07	1,04	1,69E-09	3,00	0,96	1,44	3,01	2,44	2,49
35	2,83E-07	1,05	1,35E-09	2,98	0,77	1,29	3,20	2,68	2,57
32	1,45E-07	1,06	6,42E-10	2,99	0,47	0,97	3,84	3,22	3,17
8	2,83E-07	1,05	1,26E-09	2,99	0,73	1,30	3,17	2,68	2,59
2	2,12E-07	1,05	1,04E-09	2,95	1,09	1,36	3,50	3,17	3,97
19	5,35E-07	1,04	1,52E-09	2,90	0,94	1,49	2,54	2,12	2,12
18	4,39E-07	1,04	1,60E-09	2,76	0,94	1,47	2,74	2,29	2,30
4	3,26E-07	1,04	1,60E-09	2,74	0,91	1,41	2,94	2,48	2,51
20	7,22E-07	1,04	1,35E-09	2,98	0,94	1,53	2,24	1,80	1,78
23	7,62E-07	1,03	1,31E-09	2,99	0,94	1,53	2,22	1,77	1,76
3	3,45E-07	1,03	1,92E-09	2,19	1,19	1,47	2,95	2,67	3,08
7	4,24E-07	1,08	8,52E-09	2,46	1,15	1,57	2,27	2,58	2,38
21	7,08E-07	1,04	1,43E-09	2,93	0,93	1,55	2,31	1,87	1,87
6	5,84E-07	1,08	1,30E-08	3,00	1,46	1,75	2,16	2,93	2,16
1	2,29E-07	1,04	1,06E-09	2,96	1,38	1,59	3,40	3,41	3,58
9	7,08E-07	1,03	4,21E-09	3,00	1,24	1,59	2,37	2,22	2,41
33	2,91E-07	1,03	4,84E-10	2,99	1,30	1,73	3,12	2,59	2,61
28	8,05E-07	1,01	5,04E-10	2,95	2,20	2,26	2,49	3,17	2,70
22	9,60E-07	1,03	1,12E-09	2,99	0,95	1,57	1,98	1,54	1,51
34	1,30E-06	1,02	1,64E-09	2,99	1,14	1,66	1,71	1,21	1,24
27	2,87E-06	1,00	1,13E-09	3,00	1,78	1,94	1,29	1,47	2,30
30	8,89E-12	1,42	2,18E-12	2,14	2,13	2,21	0,00	0,00	0,00
29	2,02E-07	1,01	1,29E-10	2,95	2,20	2,26	3,84	1,51	0,00

greater or equal to 0.05 (marked in white in Fig. 3). Thus, the pairs of arrhythmias for which the differences between the values of the corresponding measures are statistically significant occupy the black areas.

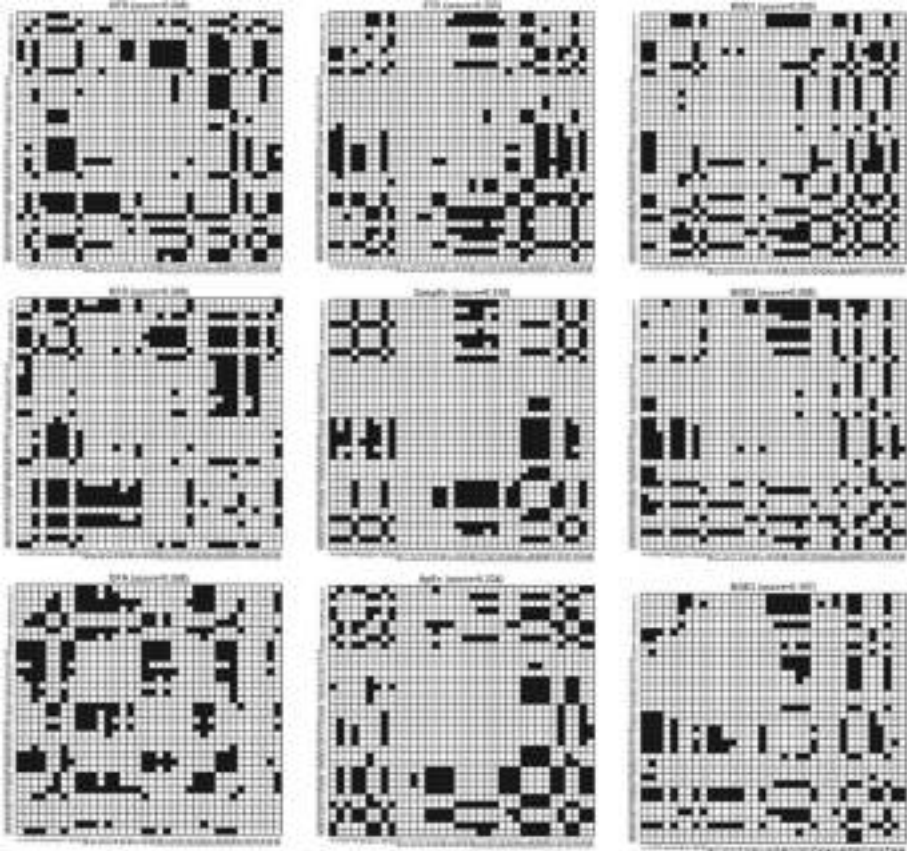


Fig. 3. The p -value matrices for nine measures. The statistically significant differences between arrhythmias for p -value < 0.05 are marked in black.

The sum of nine p -value matrices presented in Fig. 3 is shown on Fig. 4. The statistically significant differences between arrhythmias for p -value < 0.05 are marked in scale from 1 to 9 corresponding to the number of measures not allowing for the differentiation between pairs of arrhythmia. The number 9 means that the given pair is not differentiated by any of nine measures, while number 1 means that all measures, except one, allow for differentiation of this pair. For example, all measures, except *SampEn*, allow for differentiation of fine atrial fibrillation and supraventricular tachycardia (arrhythmias 2 and 9, respectively). All measures, except *MSE3*, allow for differentiation of supraventricular tachycardia (arrhythmia 9) and two other arrhythmias: nodal rhythm and 2nd degree

block (arrhythmias 8 and 35, respectively). Whereas, supraventricular tachycardia and 3rd degree block are differentiated very well by all measures, except *DFA*.

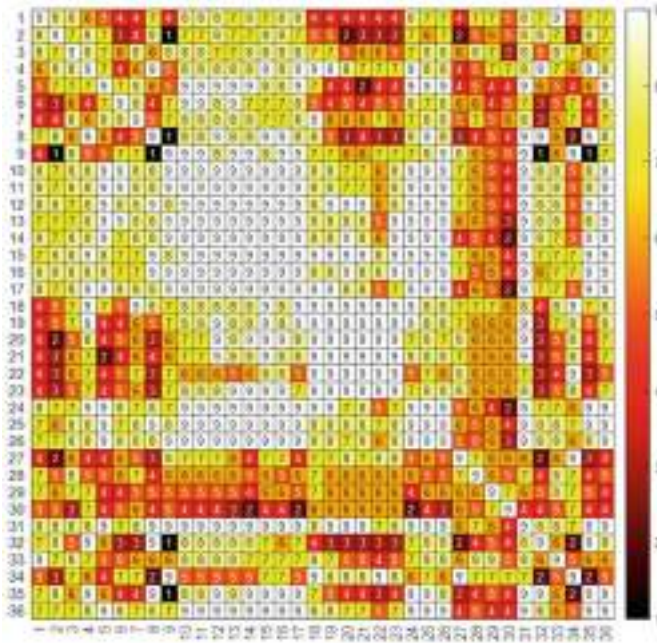


Fig. 4. The sum of p-value matrices presented in Fig. 3. The statistically significant differences between arrhythmias for p-value < 0.05 are marked in scale from 1 to 9 corresponding to the number of measures not allowing for the differentiation between pairs of arrhythmia.

The importance of features in differentiation of the regular rhythm and individual arrhythmias was summarized in Table 2.

The best features were all fractal dimension measures (*KFD*, *HFD*, *DFA*). Among all entropy measures only *ApEn* was important in the differentiation of regular rhythm and atrial tachycardia (arrhythmia 6).

To estimate an overall importance of individual features in the arrhythmias classification an efficiency score was defined as the ratio of the number of pairs for which the p-value was less than 0.05, i.e. there were statistically significant differences between both values, to the total number of pairs.

A higher score indicates higher discriminatory power of a given measure. Discrimination between arrhythmias was better for fractal features: *KFD* (score = 0.249), *HFD* (score = 0.249) and *DFA* (score = 0.248) than for variance (score = 0.233) and entropy measures: *SampEn* (score = 0.210), *ApEn* (score = 0.234), *MSE1* (score = 0.230), *MSE2* (score = 0.208) and *MSE3* (score = 0.197).

Table 2. The measures allowing for differentiation between regular rhythm and arrhythmias (N).

N	measure	N	measure	N	measure
1	<i>KFD, ApEn</i>	13	-	25	-
2	<i>KFD, DFA</i>	14	-	26	-
3	<i>DFA, HFD</i>	15	-	27	<i>KFD, HFD, ApEn, STD, MSE2</i>
4	<i>DFA</i>	16	-	28	<i>KFD, HFD, DFA, ApEn</i>
5	-	17	-	29	<i>KFD, HFD, DFA, ApEn, MSE3</i>
6	<i>ApEn</i>	18	<i>DFA</i>	30	<i>KFD, DFA, ApEn, MSE2, MSE3</i>
7	<i>DFA, ApEn</i>	19	<i>DFA</i>	31	-
8	<i>KFD</i>	20	<i>KFD, HFD</i>	32	<i>KFD</i>
9	<i>HFD, ApEn</i>	21	<i>DFA, ApEn</i>	33	<i>KFD, HFD, ApEn</i>
10	-	22	<i>KFD, HFD, ApEn, MSE2</i>	34	<i>HFD, ApEn, STD, MSE2</i>
11	-	23	<i>KFD, HFD</i>	35	-
12	-	24	-		

4 Discussion

Nonlinear dynamics methods are particularly useful for the analysis of ECG signals due to the nonlinear nature of these signals. However, most researchers did not use nonlinear measures directly to analyze ECG signals, but to analyze HRV [3–14]. In our study we proposed to apply the set of several fractal dimension and entropy measures directly to ECG signals analysis. Such approach allowed for distinction of 74% pairs of arrhythmias comparing 35 arrhythmias with each other and with normal rhythm. Moreover, most of the arrhythmias could be analyzed in very short windows of only 5 s. Only for a few arrhythmias a 30-s window was required. Whereas, the calculation of the same measures of HRV required at least a 5-min signal [12].

Our results showed that fractal dimension measures (*KFD, HFD, DFA*) are better features than entropy measures in classification of arrhythmias. Among all measures of entropy, *ApEn* was the best one. In particularly, *ApEn* was useful in distinction between regular rhythm and atrial tachycardia. Some of arrhythmias, such as missed beat, all types of premature atrial and ventricular contractions, and 1st and 2nd degree block, were indistinguishable from the regular rhythm which may be related to the fact that we analyzed the ECG signal from only one lead.

Validation of diagnostic value of these measures was performed under controlled conditions by using simulated signals from MPS450 Fluke simulator, which allows to generate a regular rhythm and a full range of 35 types of arrhythmias. This simulator is used for training purposes for physicians to recognize all possible types of arrhythmia and for testing ECG devices. Davalos and Navarrete used the same simulator to study the ventricular fibrillation [3]. In our research, we used the capabilities of this simulator to differentiate all types of arrhythmias.

In recent years, there has been a rapid development of commercial wearable ECG devices allowing for detection of arrhythmias such as tachycardia, bradycardia, and atrial fibrillation. The devices that are widely used in patient monitoring include the AliveCor KardiaMobile 6L and the Apple Watch Series 8. Moreover, the CardiacSense watch was used to detect the premature ventricular contractions (PVC) and the ventricular fibrillation. Despite these advances, commercially available devices do not offer the ability to detect all types of arrhythmia [24, 25]. Our research is heading towards the development of such device in the future. For example, the ADS1298 Analog Front-End (AFE) has a number of advantages over other components, that can allow for a technological breakthrough in ECG signal acquisition. The 24-bit resolution provides precise capture of subtle ECG changes, while the high sampling rate of up to 32 kHz supports accurate data collection even in noisy environments. In addition, low power consumption and integrated noise reduction features make it ideal for clinical applications, especially in wearable ECG devices used in real-time patient monitoring at home [24]. Combining advanced methods of signal acquisition and nonlinear analysis in an ECG device has the potential to outperform the currently available commercial devices.

One of the limitations of our study is the use of only one ECG lead. Further studies should be performed with simulated multi-channel ECG signals. Furthermore, the results should be verified using real signals recorded in patients. Moreover, no noise was added to the generated signals. However, the results depend on the quality of the ECG signals. In our previous study, we showed that HFD can differentiate ECG signals with and without motion and muscle artifacts [26]. HFD was lower for ECG with motion artifacts and higher for muscle artifacts than HFD of pure ECG.

5 Conclusion

In our work we have shown that it is possible to distinguish 20 out of 35 types of arrhythmia from the regular rhythm by using a set of only four nonlinear features (*KFD*, *HFD*, *DFA* and *ApEn*) of the full range of ECG signals in arrhythmia generated by the Fluke MPS450 simulator. In our work we tested the ADS1298, which has a number of advantages useful for measuring ECG. Combining the ADS1298 with a processor programmed to calculate nonlinear measures would enable the development of a new wearable ECG device effective in monitoring patients with arrhythmia.

Acknowledgement. This work was partially funded by the European Health and Digital Executive Agency [grant no. 101128983]. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Roth GA et al (2020) Global and regional burden of cardiovascular disease and risk factors, 1990–2019: update from the GBD 2019 Study. *J Am Coll Cardiol* 76(25):2982–3021
2. Mayo Clinic (2023) Heart arrhythmia - symptoms and causes
3. Davalos GG, Navarrete PFH (2019) Application of fractal algorithms to identify cardiovascular diseases in ECG signals. *Adv Sci Technol Eng Syst J (ASTESJ)* 5(4):173–178
4. Makowiec D, Gałaska R, Dudkowska A, Rynkiewicz A, Zwierz M (2006) Long-range dependencies in heart rate signals—revisited. *Acta Phys Pol B* 37(5):1537–1546
5. Yazawa T, Katsuyama T (2009) An extra-systole arrhythmia lowers the scaling exponent: DFA as a beneficial biomedical tool. In: *International conference on artificial intelligence and computational intelligence*, Shanghai, China, pp 601–605
6. Yazawa T (2015) *Modified Detrended Fluctuation Analysis (mDFA)*. ASME Press
7. Horie T, Burioka N, Amisaki T, Shimizu E (2018) Sample entropy in electrocardiogram during atrial fibrillation. *Yonago Acta Med* 61:049–057
8. Cervigon R, Moreno J, García-Quintanilla J, Perez-Villacastin J, Castells F (2016) Entropy at the right atrium as a predictor of atrial fibrillation recurrence outcome after pulmonary vein ablation. *Biomed Eng/Biomedizinische Technik* 61(1):29–36
9. Horie T (2013) Approximate entropy of electrocardiogram signals in atrial fibrillation. *Rinsho Byori* 61(10):893–899
10. Zhao Y, He Z (2019) Pattern recognition of multiscale entropy curve for ECG signal analysis. *Appl Mech Mater* 195–196:603–607
11. Schuckers SAC (1999) Evaluation of approximate entropy for separation of ventricular tachycardia and fibrillation. In: *Proceedings of the first joint BMES/EMBS conference. 1999 IEEE engineering in medicine and biology, 21st annual conference and the 1999 annual fall meeting of the biomedical engineering society*, Atlanta, GA, USA, vol 1, p 283
12. Jovic A, Bogunovic N (2012) Evaluating and comparing performance of feature combinations of heart rate variability measures for cardiac rhythm classification. *Biomed Signal Process Control* 7(3):245–255
13. Acharya UR, Faust O, Sree SV, Swapna G, Martis RJ, Suri JS (2014) Linear and nonlinear analysis of normal and CAD-affected heart rate signals. *Comput Methods Programs Biomed* 113(1):55–68
14. Gupta P, Sharma S, Anand RS, Gupta RS (2020) A novel approach for the classification of arrhythmia by using deep learning methods. *Biomed Signal Process Control* 62:102074
15. Chen C, da Silva B, Ma C, Li J, Liu C (2024) Fast sample entropy atrial fibrillation analysis towards wearable device. In: Wang G, Yao D, Gu Z, Peng Y, Tong S, Liu C (eds.) *12th Asian-Pacific conference on medical and biological engineering. APCMBE 2023. IFMBE Proceedings*, vol 103. Springer, Cham
16. Texas Instruments (2015) ADS129x Low-Power, 8-Channel, 24-Bit Analog Front-End for Biopotential Measurements (Rev. K)
17. Fluke Biomedical (2005) MPS450 Patient Simulator
18. Higuchi T (1988) Approach to an irregular time series on the basis of the fractal theory. *Physica D* 31(2):277–283
19. Katz MJ (1988) Fractals and the analysis of waveforms. *Comput Biol Med* 18(3):145–156
20. Peng CK, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos Interdisc J Nonlinear Sci* 5(1):82–87
21. Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol-Heart Circulatory Physiol* 278(6):H2039–H2049

22. Pincus SM (1991) Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci* 88(6):2297–2301
23. Costa M, Goldberger AL, Peng CK (2002) Multiscale entropy analysis of complex physiologic time series. *Phys Rev Lett* 89(6):068102
24. Prieto-Avalos G, Cruz-Ramos NA, Alor-Hernández G, Zavala-Galindo G (2022) Wearable devices for physical monitoring of heart: a review. *Biosensors* 12(5):292
25. Dahiya ES, Kalra AM, Lowe A, Anand G (2023) Wearable technology for monitoring electrocardiograms (ECGs) in adults: a scoping review. *Sensors* 24(4):1318
26. Olejarczyk E, Raus-Jarzabek E, Massaroni C (2024) Automatic identification of movement and muscle artifacts in ECG based on statistical and nonlinear measures. In 2024 IEEE international workshop on metrology for industry 4.0 & IoT (MetroInd4.0 & IoT), Firenze, Italy, pp 269–273. <https://doi.org/10.1109/MetroInd4.0IoT61288.2024.10584239>



QRS Detection Based on the Matching Pursuit Algorithm

Mohammed Owahidur Rahman¹, Piotr Augustyniak¹, and Elzbieta Olejarczyk^{1,2}(✉)

¹ Department of Biocybernetics and Biomedical Engineering, AGH University of Krakow,
30-059 Krakow, Poland
eolejarczyk@agh.edu.pl

² Nalecz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences,
02-109 Warsaw, Poland

Abstract. The purpose of this study was the detection of QRSs in ECG signals using the Matching Pursuit (MP) method. Five groups of patients were considered: healthy controls (NORM), patients with ST-T change (STTC), myocardial infarction (MI), conduction disturbance (CD), and hypertrophy (HYP). The ECG signals were decomposed using MP algorithm in set of functions, so called atoms, characterized by specific energy, amplitude, frequency and scale, corresponding to different ECG patterns, QRS complexes and T waves. A threshold for frequency ranging from 6 Hz to 45 Hz and scale value from 0.02 to 0.069 was applied to identify the QRS patterns. The proposed algorithm allowed for the identification of QRSs with a high performance independently on the patient group. The average values of sensitivity, specificity, and positive and negative predictive values were 98.94%, 100%, 100%, and 99.30%, respectively. The QRSs detection using an algorithm based on MP decomposition and on the application of two thresholds for frequency and scale is very efficient, and it can be applied in ECG wearable devices for monitoring patients in real time.

Keywords: ECG · QRS detection · Matching Pursuit method

1 Introduction

Electrocardiography is a very important tool in the diagnosis of heart disease. The electrocardiogram (ECG) is characterized by a repeating sequence of P, QRS, and T waves associated with each beat (Fig. 1). QRS complexes represent a ventricular depolarization. Thus, their detection is a fundamental task in ECG signal processing. Various techniques have been developed over the years, ranging from classical rule-based algorithms to advanced deep learning approaches. One of the most widely used classical methods is the Pan-Tompkins algorithm, which applies bandpass filtering, differentiation, squaring, and a moving window integration to enhance the QRS complex and detect R-peaks [1]. An improved version of this method, known as the Hamilton-Tompkins algorithm, introduced additional rules for dealing with noise, arrhythmia, and sudden heart rate changes [2]. Recently, a new QRS detection algorithm has been developed, which is a modification of the standard Pan-Tompkins algorithm [3]. It consists in normalizing ECG signals

in short windows (1200 ms). An algorithm based on Wavelet Transform, consisted in the wavelet decomposition of ECG signal into multiple frequency bands, was introduced to improve the robustness of QRS detection to noise [4]. Similarly, the methods based on Hilbert Transform employ signal envelope extraction to improve QRS detection accuracy, especially in real-time applications [5]. With the development of artificial intelligence, the approaches based on machine learning have been explored, including Support Vector Machines (SVM), Artificial Neural Networks (ANN), and deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). These models can learn from large ECG datasets to generalize across different patient conditions, improving accuracy and robustness [6]. The QRS detection based on deep learning models like CNNs and Long Short-Term Memory (LSTM) networks has outperformed traditional methods, achieving accuracy similar to that of a cardiologist in arrhythmia detection and QRS segmentation [7]. Another efficient approach is adaptive thresholding and moving average, which optimizes detection for low-power embedded systems by applying adaptive logic combined with a moving average filter to identify QRS complexes efficiently [8].

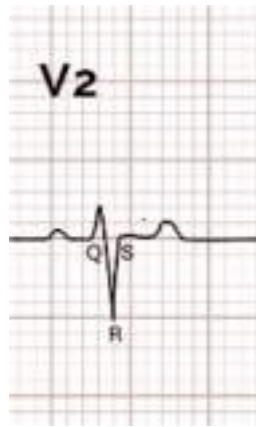


Fig. 1. A typical ECG pattern with V2 lead.

Each method has its advantages and limitations. While classical techniques like Pan-Tompkins and Hamilton-Tompkins are computationally efficient and suitable for real-time processing, they may struggle with noisy signals. Wavelet and Hilbert transform-based methods improve performance in non-ideal conditions but can be computationally expensive. Machine learning and deep learning approaches offer superior accuracy and adaptability but require extensive training data and higher computational resources.

The Matching Pursuit (MP) algorithm based on the decomposition of ECG signal into its main waveform components can also be useful for the detection of QRS complexes [9]. In this study we proposed to extend the application of the MP method for the detection of QRSs using atom features such as time position, amplitude, energy, frequency, phase and scale with appropriate thresholds.

2 Materials and Methods

2.1 ECG Recording and Preprocessing

The publicly available PTB-XL ECG database of 10-s 12-lead ECG signals from 3679 patients were used in this study [10]. Five different classes: 1048 healthy controls normal (NORM) consist, 1048 patients with conduction disturbance (CD), 535 with hypertrophy (HYP), 1048 with ST-T change (STTC) and 1048 with myocardial infarction (MI). The signals were recorded with an amplitude of the ECG is 0.5–5 mV and a sampling frequency of 100 Hz. The V2 lead was chosen for further analysis. The reason for selecting this location is its excellent view of the anterior and septal regions of the left ventricle. The left anterior descending (LAD) coronary artery, which is one of the main coronary arteries, provides the majority of the supply to these areas.

2.2 Matching Pursuit Method

The feature which distinguishes Matching Pursuit (MP) method from other methods, is a great variety of functions in the dictionary [11, 12]. In practice, Gabor functions (Gaussian modulated sinusoids) are used as basic waveforms due to their optimal time-frequency localization. To this set of functions, so called atoms, sinus and cosinus functions as well as Dirac's delta functions can be added.

A waveform from a time-frequency dictionary can be expressed as a translation (u), dilation (s) and modulation (ω) of a window function $g(t)$:

$$g_{\gamma}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) \cdot e^{i\omega t} \quad (1)$$

The MP method decomposes ECG signal in set of atoms. The atom table containing the values of time position, amplitude, energy, frequency, phase, and scale is generated. T waves and QRS complexes are reconstructed by summing the appropriate Gabor atoms. In this study, the MP method was applied to the ECG data using SVAROG software [13, 14] to recover the atom table.

The reconstruction of T waves and QRS complexes consisted in summing all Gabor functions using the parameters contained in the atom table. At the beginning, an empty array of length corresponding to the total number of samples in the signal was created.

A single atom is described by the following Gabor function:

$$g(t) = A \cdot \exp\left(-\frac{(t-t_0)^2}{2s^2}\right) \cdot \cos(2\pi f(t-t_0) + \varphi) \quad (2)$$

Here, A represents the amplitude, t_0 - time position of atom, s - scale, f - frequency, and φ - phase shift.

The final reconstructed signal is the sum of all individual Gabor functions:

$$S(t) = \sum g_i(t) \text{ for } i = 1, \dots, N \quad (3)$$

The MP calculation was performed for a joint signal from all patients in each group. The combined signal was decomposed into atoms using the MP algorithm. Then, the QRSs were identified by applying two thresholds for frequency and scale. They were chosen based on the histograms of these parameters extracted from atom tables.

2.3 Statistical Analysis

The performance of the QRS complexes detection was tested with the databases of all patients of each class and evaluated using the standard metrics, such as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) according to the following formulas:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$PPV = \frac{TP}{TP + FP} \quad (6)$$

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

where TP – QRSs marked manually in the original signal and QRSs reconstructed from the atom table using two thresholds; TN – remaining atoms not corresponding to QRSs marked manually and identified automatically; FP – no QRSs marked manually but identified automatically; FN – QRSs marked manually but no identified automatically.

3 Results

3.1 Identification of QRSs

The distributions of frequencies and scales from atom table for all patients from five groups are presented in Fig. 2. The thresholds for frequency from 6 to 45 Hz and for scale from 0.02 to 0.069 were chosen for identification of QRSs. These ranges are different than ranges for other ECG components independently of the patient condition.

3.2 ECG Analysis Using Matching Pursuit Method

Examples of original and reconstructed ECG signals and reconstructed QRSs obtained as a result of decomposition of two 10-s ECG signals in NORM and MI are presented in Fig. 3.

The performance results of QRS detection in all patients separately for each group of patients are presented in Table 1. The MP algorithm demonstrated a high effectiveness regardless of the patient group. The average sensitivity, specificity, and positive and negative predictive values were 98.94%, 100%, 100%, and 99.30%, respectively.

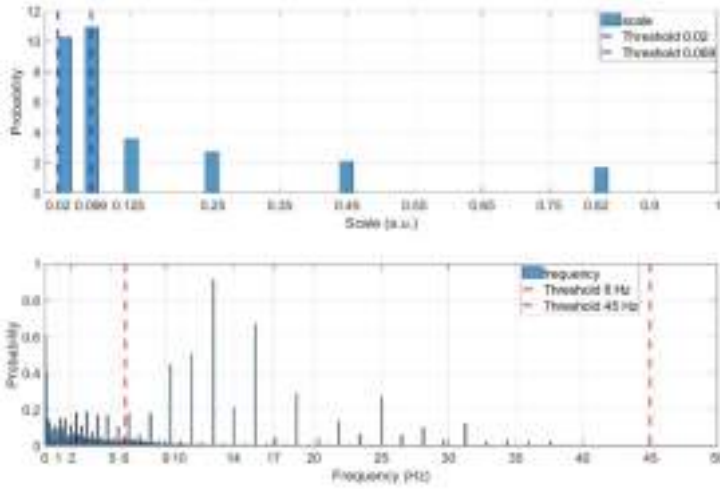


Fig. 2. The distributions of frequencies and scales from atom table for all patients. The QRSs were identified for thresholds marked with the vertical dashed lines.

4 Discussion and Conclusions

The Matching Pursuit method is a powerful technique of the ECG signal analysis allowing for an efficient identification of ECG patterns and determination of parameters of the function describing their shape. The choice of a proper dictionary is essential to gather these features of the signal that are useful for the QRS detection [15, 16]. The performance of the algorithm depends on an appropriate dictionary design, computational complexity, and robustness to extreme variability or noise.

Recently, an optimized implementation of the MP algorithm, combining high computational efficiency with robust and accurate decomposition capabilities, was presented [17]. It is particularly well-suited for applications requiring precise parameter estimation, such as ECG signal analysis.

In our study, the obtaining of the high performance of the MP algorithm in the QRS detection was possible by application of the efficient *empi* algorithm [17] and by the choice of appropriate thresholds for frequency and scale values derived from the atom table provided by this algorithm. The robustness of the proposed method, even when applied to datasets with varying signal quality and patient conditions, emphasizes its reliability and clinical potential. The morphology of ECG and its components, such as QRS and T waves, depends on the patient's condition. In particular, changes in the morphology of T waves are observed in ST-elevation myocardial infarction (STEMI). This issue will be a subject of our future study. For the QRS detection, a full range of parameters (frequency and scale) for both healthy and pathological conditions was considered. These ranges are separated from the ranges of these parameters for T waves, regardless of the patient's condition, which affects the accuracy of QRS detection. Thus, this algorithm can be useful in ECG wearable devices for monitoring patient's cardiac activity in real time.

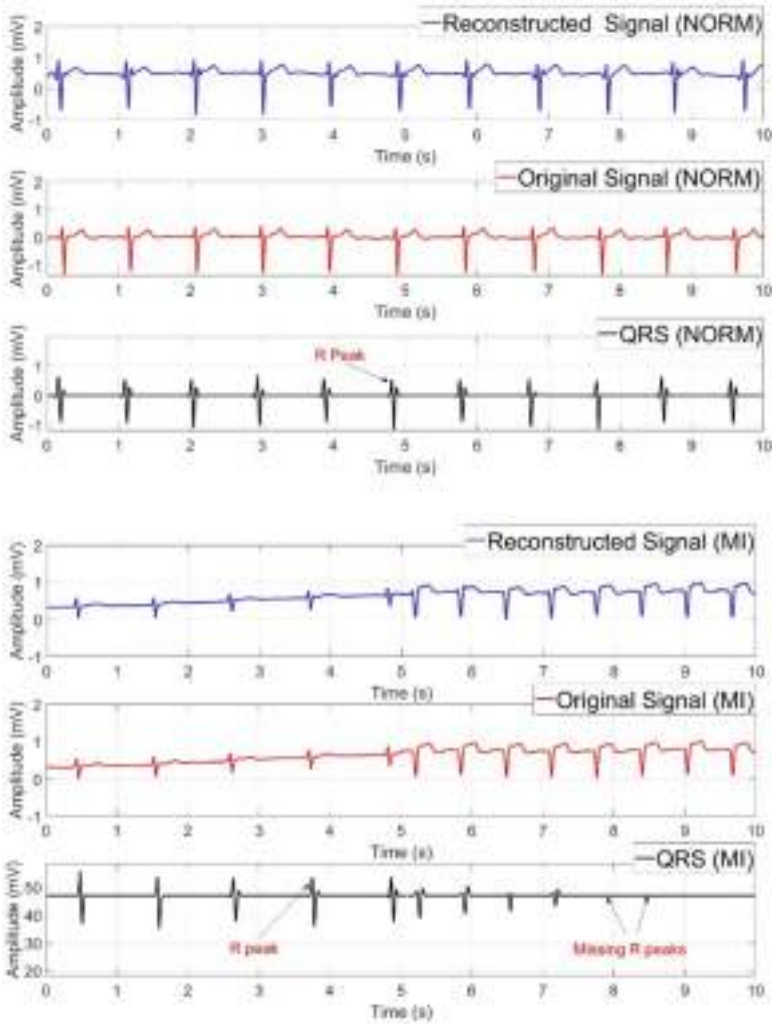


Fig. 3. Examples of original and reconstructed ECG signals and reconstructed QRS signal in NORM (A) and in MI (B).

However, MP methods are not without their limitations. In real-time applications, the computational burden of exhaustive dictionary searches and the algorithm's sensitivity to the initial dictionary design can cause challenges. If the dictionary does not adequately represent the signal's wide variability, the MP method may become less effective in the presence of severe noise or artifacts. Future research should be focused on addressing these challenges. Integrating machine learning techniques, like deep neural networks, with MP could lead to the creation of adaptive, real-time dictionary learning frameworks that improve the reliability and speed of ECG analysis. Furthermore, further studies are needed to evaluate the clinical efficacy of MP-based systems in large-scale trials and to

Table 1. Performance results of QRSs detection in five groups of patients.

Class	Sensitivity	Specificity	PPV	NPV	TP	FP	FN	TN
NORM	99.26	100	100	99.46	8662	0	64	11840
MI	99.40	100	100	99.61	7694	0	45	11731
STTC	99.82	100	100	99.85	9361	0	16	10720
CD	98.47	100	100	99.12	7444	0	115	13034
HYP	97.77	100	100	98.44	8633	0	198	12559

explore their integration into wearable monitoring devices. Moreover, the MP algorithm can be applied also for detection and analysis of other ECG components such as T waves, what will be the subject of our future work.

Acknowledgement. This work was partially funded by the European project PROVIDE (“Prediction and prevention of cardiovascular diseases in pre- and type 2 diabetes”, Grant Agreement no. 101128983, call topic EU4H-2022-PJ-1) and by AGH University of Krakow in 2024. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

References

1. Pan J, Tompkins WJ (1985) A real-time QRS detection algorithm. *IEEE Trans Biomed Eng* 32(3):230–236
2. Hamilton PS, Tompkins WJ (1986) Adaptive and real-time QRS detection using a modified lead II ECG. *IEEE Trans Biomed Eng* 33(12):1157–1165
3. Rahman MO, Augustyniak P, Olejarczyk E (2024) The QRS detection using the Modified Pan-Tompkins algorithm. In: 2024 IEEE international workshop on metrology for industry 4.0 & IoT (MetroInd4.0 & IoT), 29–31 May 2024, Firenze, Italy
4. Kadambe S, Murray RG, Boudreaux-Bartels GF (1999) Wavelet transform-based QRS complex detector. *IEEE Trans Biomed Eng* 46(7):838–848
5. Arzeno AJ, Deng ZD, Poon CS (2008) Analysis of first-derivative based QRS detection algorithms. *IEEE Trans Biomed Eng* 55(2):478–484
6. Sayadi O, Shamsollahi MB (2010) ECG denoising and QRS detection using a nonlinear Bayesian filtering framework. *IEEE Trans Biomed Eng* 57(4):826–837
7. Hannun AY et al (2019) Cardiologist-level arrhythmia detection with deep neural networks. *Nat Med* 25(1):65–69
8. Elgendi M (2013) Fast QRS detection with an optimized knowledge-based method: evaluation on 11 standard ECG databases. *PLoS ONE* 8(9):e73557
9. Jadvar H, Jenkins J, Arzbaecher R (1988) Computer ST segment measurement during trans-esophageal atrial pacing stress. In: *Proceedings of computers in cardiology*, Washington, DC, USA, pp 227–229
10. Wagner P, Strodthoff N, Bousseljot R, Samek W, Schaeffter T (2022) PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). *PhysioNet*. <https://doi.org/10.13026/kfzx-aw45>

11. Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans Signal Process* 41(12) (1993)
12. Blinowska KJ, Durka PJ (2001) Unbiased high resolution method of EEG analysis in time-frequency space. *Acta Neurobiol Exp* 61:157–174
13. Kus R, Rozanski P, Durka P (2013) Multivariate matching pursuit in optimal Gabor dictionaries: theory and software with interface for EEG/MEG via Svarog. *Biomed Eng Online* 12:94
14. Rozanski P (2020) Effects of envelope and dictionary structure on the performance of matching pursuit. *IET Signal Process* 14(2):89–96
15. Tcheou MP, Lovisolo L, da Silva EA, Rodrigues MA, Diniz PS (2007) Optimum rate-distortion dictionary selection for compression of atomic decompositions of electric disturbance signals. *IEEE Signal Process Lett* 14(2):81–84
16. Yaghoobi M, Daudet L, Davies ME (2009) Parametric dictionary design for sparse coding. *IEEE Trans Signal Process* 57(12):4800–4810
17. Rozanski PT (2024) Empi: GPU-accelerated matching pursuit with continuous dictionaries. *ACM Trans Math Softw* 50(3):17



Influence of the 3D Printing Infill, Pattern and CT kV on the Hounsfield Units for Biomedical Applications

Nikolay Dukov¹ , Kristina Bliznakova¹ , Yanka Baneva^{1,2} ,
and Zhivko Bliznakov¹ 

¹ Medical University of Varna, Varna, Bulgaria
{ntdukov, kristina.bliznakova, yanka.baneva,
zhivko.bliznakov}@mu-varna.bg

² St. Marina University Hospital, Varna, Bulgaria

Abstract. Polylactic Acid (PLA) is a biodegradable thermoplastic polymer, commonly used with low-cost fused filament fabrication (FFF) 3D printers. Amongst the many biomedical applications of PLA, is the use in the Radiology field, where it is exploited for preparing physical phantoms. The aim of this study is to investigate the influence of 3D printing patterns on the Hounsfield Units (HU) of Computed Tomography (CT) scanned samples. For this study, we fabricated samples with two different infill patterns—cubic subdivision and quarter cubic—across a range of infill densities (10% to 100%). The samples were scanned at a CT facility using three different kV settings. Regions of interest (ROIs) were selected from the CT images of the scanned samples, and the corresponding HU were measured. The results demonstrated that both infill pattern and density significantly influence the HU values. Both patterns were found to be suited for representing lung tissue, while the Quarter Cubic pattern was appropriate for replicating the radiological properties of adipose tissue, as well. The kV settings had a slight effect on the HU measurements. These findings will be applied to the development of anthropomorphic phantoms for radiological applications.

Keywords: 3D printing · PLA · printing pattern and infill · Hounsfield units · physical phantoms

1 Introduction

Polylactic Acid (PLA) is a biodegradable thermoplastic polymer, very popular with the low-cost FFF-based (fused filament fabrication) 3D printers [1]. It is built from renewable resources such as corn starch, sugarcane, or tapioca roots [2]. Because PLA possesses biocompatibility and biodegradability, it finds many applications for biomedical applications, such as in prototyping of medical devices and implants, tissue engineering and scaffolds, drug delivery systems as well as structures and surgical implants, biodegradable stents, dental applications [2, 3].

Among the biomedical applications of PLA, its use in the fields of radiology and radiotherapy stands out. One key application is the creation of physical test objects, or phantoms, which are used for quality control, testing and optimizing clinical protocols, and evaluating new X-ray imaging devices. In these cases, biocompatibility is not a primary concern. Instead, the most critical factor is that the X-ray absorption coefficients of PLA closely resemble those of human tissues, ensuring accurate simulation and testing outcomes.

Studies have shown that PLA material has a good agreement with the human organs since all of them consider as a soft tissue in terms of mass attenuation coefficients in the diagnostic range [4, 5]. In the radiotherapy range, other experimental studies showed that it is very appropriate to replicate the human tissues [5].

In Computed Tomography (CT) scans, the Hounsfield Unit (HU) represents the degree of X-ray attenuation by the scanned object. FFF 3D printing, however, uses filaments to create objects with varying infill factors, ranging from 0% to 100%. For example, Assemany *et al.* [6] utilized 90% infill density PLA filament to evaluate its suitability for dose assessment in pediatric and adolescent patients undergoing cone-beam CT exams. At this infill density, the HU values ranged from 98 HU to 269 HU across a voltage range of 100 kV to 150 kV. The study concluded that PLA demonstrated superior results compared to the reference PMMA material, making it a promising candidate for such applications.

Studies have shown that using a lower infill factor results in reduced HU, effectively simulating the radiological properties of human tissues with lower electron densities. For example, Dancewicz *et al.* [5] demonstrated this effect with PLA infill densities of 10%, 30%, 50%, 70%, and 90%, scanned at 80–120 kV. The results indicated that 30% PLA can simulate lung tissue, 50% PLA can replicate adipose tissue, while 70% PLA and 90% PLA are suitable for modeling brain and bone tissues, respectively. This approach has been adopted by a limited number of researchers to create various organs using 3D printers with different infill densities, enabling more accurate tissue-mimicking models [7–9].

Investigations related to how different infill and printing patterns influence the HU in CT imaging are limited, but important. One of the study is reported by Talalwa *et al.* [10] for different infill densities and patterns, showing that PORO-LAY materials are capable in achieving low and high values of HU ranges in the range -990 to $+950$. The results also indicate that the shape of the infill pattern influences the HU values, with percentage differences ranging from 3% to 33%, depending on the selected infill structure. Similar results are reported for ABS material, showing that modification of the filling form as well as its percentage significantly influenced the HU values [11].

The aim of this study is to investigate the influence of 3D printing patterns on the HUs of the scanned samples printed with PLA.

2 Materials and Methods

To achieve the aim of this study, we designed a cube, which was loaded into 3D printing slicing software and assigned specific settings, resulting in objects with varying characteristics.

2.1 Sample

A computer model of a cube with a size of $2 \times 2 \times 2$ cm was generated using the free software OpenSCAD (<https://openscad.org>, see Fig. 1). The computer model was later exported as an STL file in order to be 3D printed.

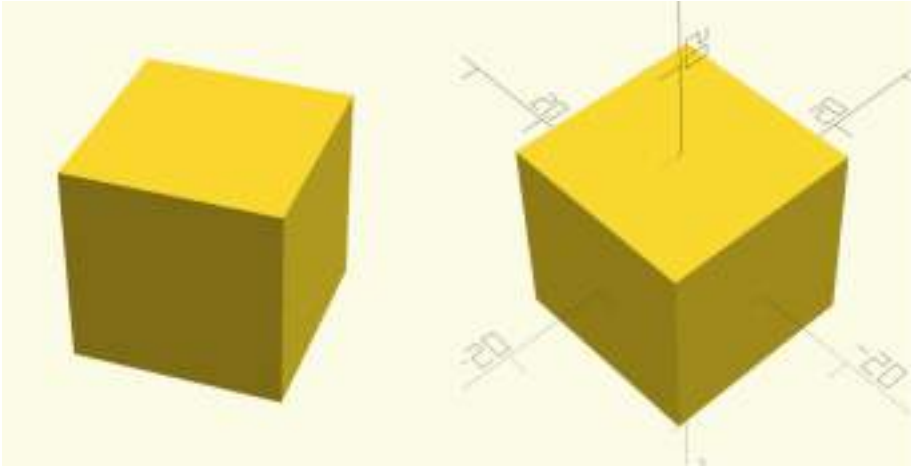


Fig. 1. Screenshots of the generated cube with dimensions of $2 \times 2 \times 2$ cm, in OpenSCAD (<https://openscad.org>).

2.2 Printer, Filling Pattern and Materials

For the 3D printing of multiple cubes, the free slicing software Ultimaker Cura (<https://ultimaker.com/software/ultimaker-cura/>) was used. The loaded STL file of the cube was duplicated to produce 20 cubes. To examine the impact of the infill pattern on the HUs, several default printing settings were adjusted. The top and bottom layers, as well as the walls, were set to zero (removed), ensuring that only the infill pattern was printed. Additionally, the layer height was set to 0.2 mm. For ten of the cubes, the Cubic Subdivision infill pattern was selected, while the remaining ten cubes were printed with the Quarter Cubic infill pattern. Each cube in both patterns was assigned a specific infill percentage, starting at 10% and increasing by 10% increments up to 100%, resulting in ten cubes per pattern with identically increasing infill (see Fig. 2).

We used a budget-friendly hobbyist FFF 3D printer (Longer LK4, <https://longer3d.com/>) with a 0.4 mm brass nozzle for the 3D printing of the cubes and utilizing a generic low-cost PLA filament. The printing temperatures and speed were set according to the default profile of the printer in Ultimaker Cura. More specifically, the printing temperature was set to 200 °C, temperature of the printing bed was set to 60 °C and for printing speed – 50 mm/s.

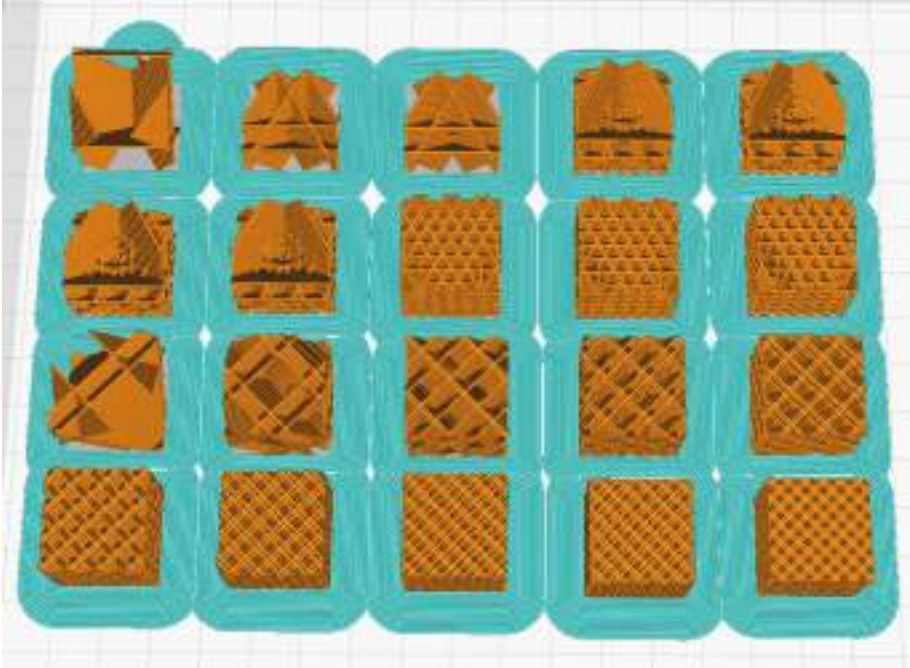


Fig. 2. A screenshot of the sliced cubes in Ultimaker Cura is shown. The first two rows represent the Cubic Subdivision infill pattern, while the third and fourth rows correspond to the Quarter Cubic infill pattern. For each pattern, the infill begins at 10% and increases in 10% increments, reaching up to 100%.

2.3 Experimental Protocol

The samples were scanned using a Siemens SOMATOM Force CT scanner, at three energy levels: 80 kV, 100 kV, and 120 kV, with a slice thickness of 1 mm. This resulted in an image dataset with a voxel size of $0.2 \times 0.2 \times 1$ mm.

2.4 Measurements

Measurements were conducted on 20 consecutive slices in the axial view. For each cube an identical square region of interest (ROI) was used with a size of $13 \text{ mm} \times 13 \text{ mm}$. An example of the ROI taken for one slice can be seen on Fig. 3. The mean and standard deviation were calculated for each cube based on the obtained results. All measurements were performed using ImageJ [12].

3 Results

The 3D-printed cubes produced with the two patterns are shown in Fig. 4. The first two rows correspond to the Cubic Subdivision pattern. In the first row, the leftmost cube has an infill of 10%, with each subsequent cube increasing the infill by 10%. This progression



Fig. 3. An axial slice of the scanned cubes showing a selected ROI (yellow square) for one of the cubes. The same ROI size and positioning were consistently applied across twenty consecutive slices of each cube.

results in the rightmost cube of the first row having an infill of 50%. The second row continues this sequence, starting with a cube at 60% infill and ending with a cube at 100% infill.

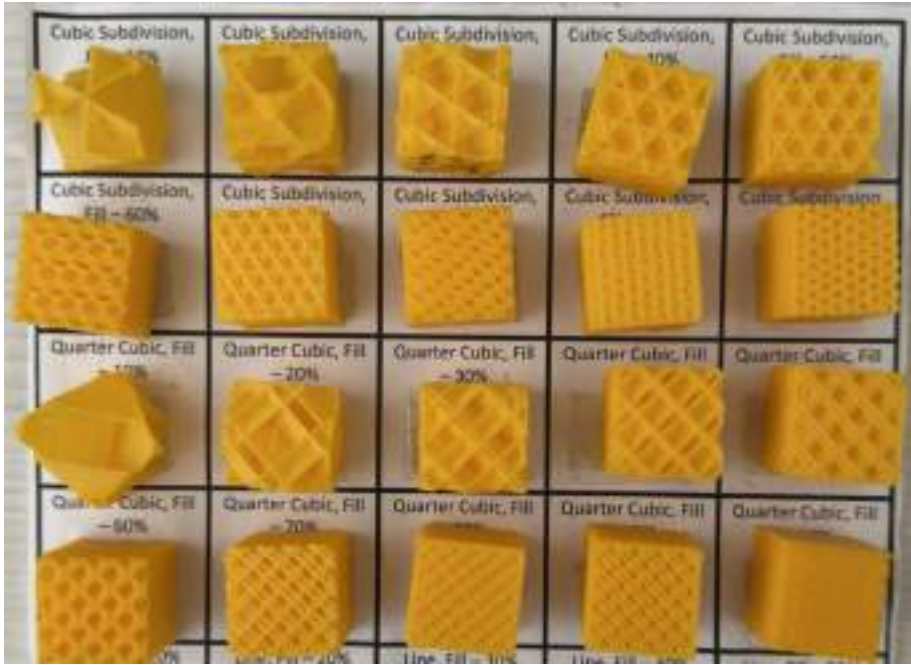


Fig. 4. Resulting printed cubes with the examined infill patterns – Cubic Subdivision and Quarter Cubic. The infill of the cubes varies from 10% to 100%, with an incrementing step of 10%.

The third and fourth rows of Fig. 4 display the cubes printed with the Quarter Cubic infill pattern. The arrangement of these cubes (infill percentage sequence) follows the same progression as that of the Cubic Subdivision pattern.

A representative scanned image of all the studied cubes is shown in Fig. 5. The positions of the cubes corresponds to the position and patterns of the cubes shown in Fig. 2 and Fig. 4.



Fig. 5. Screenshot of a coronal slice (a middle slice) of the scanned cubes at 80 kV.

The results of the HU measurements are graphically presented in Fig. 7 and Fig. 8 for the Cubic Subdivision and Quarter Cubic infill patterns, respectively.

For the Cubic Subdivision pattern (Fig. 7), it can be observed that the HUs increase linearly from 10% to 60% infill. However, at 70% infill, there is a sudden drop in the measured HUs, falling below the values measured at 60%. A similar behavior is noted at 90% infill, although the drop in this case is less pronounced.

The source of this behavior can be identified in Fig. 5. The second row from the top shows the cubes printed with the Cubic Subdivision pattern, ranging from 60% to 100% infill (left to right). From 70% to 100% infill, hollow regions are visible inside the cubes, primarily in their central areas. This phenomenon is observed only at higher infill settings. It appears that for this pattern, the outer layers of the cubes are denser, while the inner sections remain sparse.

In the context of 3D printing, this design has practical advantages. It allows for structural strength in the outer walls, for cases where it is most needed, while reducing material usage in the supposed less critical inner regions. Compared to a fully dense structure, this approach offers lower material usage. A visual inspection of the sliced computer model with the Cubic Subdivision pattern clearly shows the concentration of infill at the edges of the cube at the expense of the inner parts, as seen from Fig. 6.

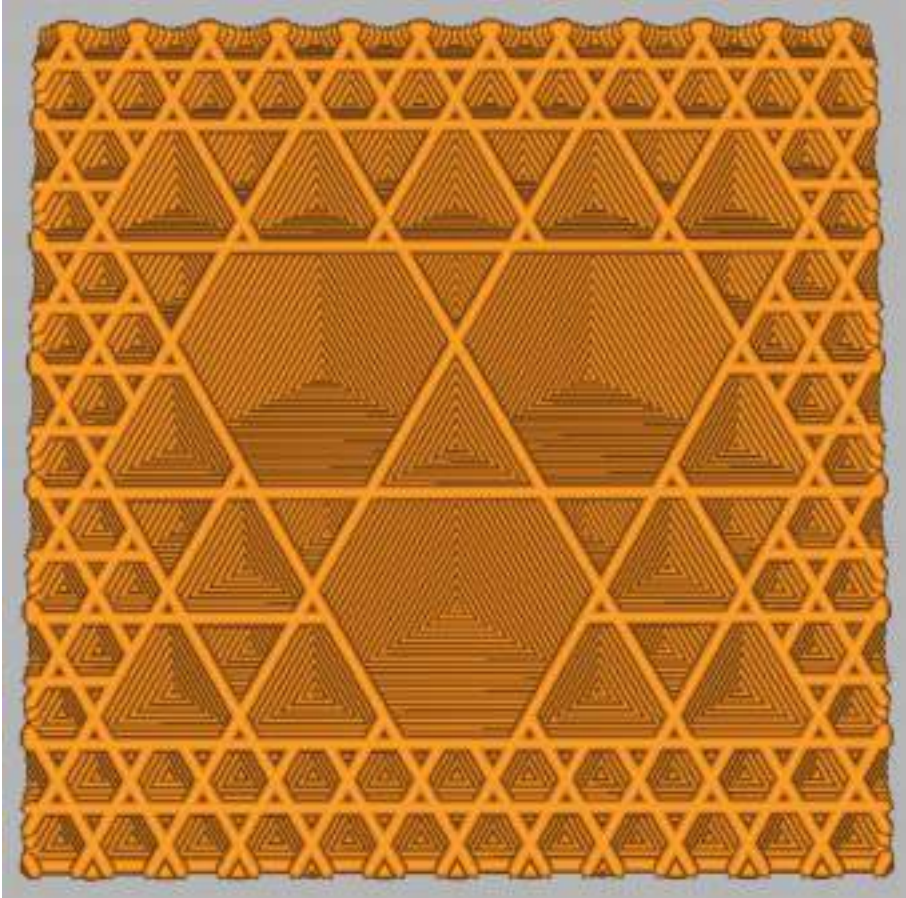


Fig. 6. Screenshot of the sliced computer model of a cube with Cubic Subdivision pattern and 100% infill. The image shows a section at the middle (at 1 cm height) of the cube. The sparsity of the infill in the inner part of the cube is visible.

Additionally, the HU values for a given infill percentage within the same pattern remain consistent across the three CT energy levels (80 kV, 100 kV, and 120 kV), as demonstrated in Fig. 7. This aligns with the observations reported in the experimental study conducted by Ma *et al.* [13].

The calculated standard deviation for HUs of the Cubic Subdivision pattern ranges from ± 15 HUs to ± 25 HUs for infills of 10% to 60%, increases to between ± 70 HUs to ± 90 HUs for 70% and 80% infills and further rises to ± 125 HUs to ± 140 HUs for 90% and 100% infills. These observations are consistent across all three energy levels used in the study.

The Quarter Cubic infill pattern (Fig. 8) exhibits linear behavior with respect to changes in infill percentage – an increase in the infill percentage results in a positive increase of the HUs. Moreover, from Fig. 8 it is noticeable that for this pattern when

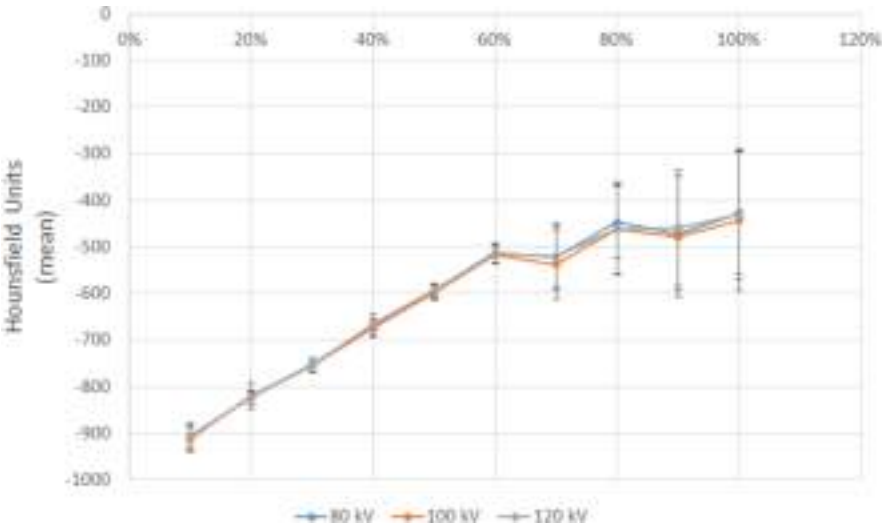


Fig. 7. Changes of the Hounsfield Units depending on the infill percentage for pattern Cubic Subdivision, for the three energies used – 80 kV, 100 kV and 120 kV.

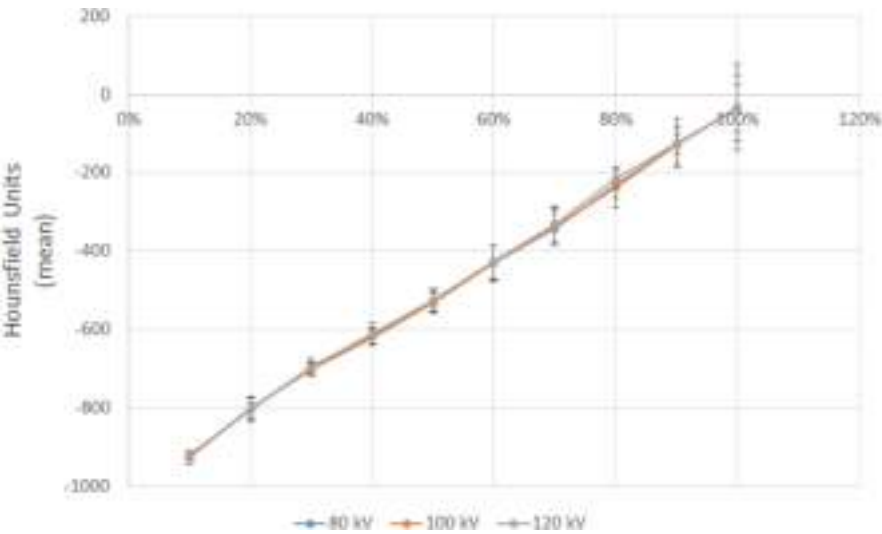


Fig. 8. Changes of the Hounsfield Units depending on the infill percentage for pattern Quarter Cubic, for the three energies used – 80 kV, 100 kV and 120 kV.

the infill is increased with 10%, the change in the HUs values is around 100 HU. This predictable relationship enhances the utility of the pattern, making it straightforward to apply in specific use cases. Similarly to the Cubic Subdivision pattern, the HU values for a given infill percentage of the Quarter Cubic pattern remain consistent across the three CT energy levels (80 kV, 100 kV, and 120 kV).

Obtained HU ranges reveal that both infill patterns can be used to approximate the radiological appearance of lung tissue with typical values of -785.7 ± 11.5 [14]. The Quarter Cubic infill pattern can be used to replicate the adipose tissue, since the later approximately has values in the range $[-120 \text{ HU}: -60 \text{ HU}]$ at 70 kV to $[-90 \text{ HU}: -20 \text{ HU}]$ at 120 kV [15].

The calculated standard deviation of the HU values for the Quarter Cubic pattern ranges from ± 15 HUs to ± 25 HUs for infills of 10% to 50% and increasing to between ± 40 HUs to ± 50 HUs for 60% and 90% infills. However, at 100% infill, a notable difference in the standard deviation in relation to the energy used is observed. For 80 kV the standard deviation is ± 60 HU, for 100 kV the calculated standard deviation increases to ± 83 HU and at 120 kV it further increases to ± 110 HU.

Limitations of this study include the use of a single printing material, namely PLA, and a restricted selection of infill patterns. These constraints limit the ability to simulate higher-density tissues. Future work will address these limitations by exploring additional infill patterns. Expanding to other materials, although envisioned, will likely introduce challenges, particularly when printing more complex geometries. PLA is arguably among the easiest materials to handle, offering reliable printability and reproducibility, across different 3D printers – a consistency that may be more difficult to achieve with alternative printing materials.

4 Conclusions

We looked at two infill printing patterns and the effects of the infill percentage on the HU values at three different energies (80 kV, 100 kV and 120 kV). It was found that the patterns and the infill percentage corresponding to a given pattern exhibit specific behavior. Both patterns were found to be well suited for representing lung tissue, while the Quarter Cubic pattern was also appropriate for replicating adipose tissue. It was noticed that the kV settings had minimal effect on the HU values of the studied objects. These findings will be applied to the development of an anthropomorphic phantoms for radiological applications.

Further investigations into the effects of patterns and infill percentages on the HUs are underway, along with the development of custom patterns tailored to specific HU values.

Acknowledgement. This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0009-C02. The work is performed by the ELPIDA 3.1.5 research group, which is part of the MUVE-TEAM project No BG-RRP-2.004-0009-C02.

Conflict of Interest. The authors declare that they have no conflict of interest.

References

1. Vallejos Baier R et al (2021) Structure-function assessment of 3D-printed porous scaffolds by a low-cost/open source fused filament fabrication printer. *Mater Sci Eng C Mater Biol Appl* 123:111945

2. Moya-Lopez C et al (2022) Polylactide perspectives in biomedicine: from novel synthesis to the application performance. *Pharmaceutics* 14(8) (2022)
3. Bliznakova K, Milev M, Dukov N, Atanasova V, Yordanova M, Bliznakov Z (2024) Pilot study on the development and integration of anthropomorphic models within the dental technician curriculum. *Dent J-Basel* 12(4) (2024)
4. Alssabbagh M, Tajuddin A, Manap MB, Zainon R (2017) Evaluation of nine 3D printing materials as tissue equivalent materials in terms of mass attenuation coefficient and mass density. *Int J Adv Appl Sci* 4(9):168–173
5. Dancewicz OL, Sylvander SR, Markwell TS, Crowe SB, Trapp JV (2017) Radiological properties of 3D printed materials in kilovoltage and megavoltage photon beams. *Phys Medica* 38:111–118
6. Assemany LPF, Rodrigues O, da Silva E, Potiens MDA (2020) Evaluation of 3D printing filaments for construction of a pediatric phantom for dosimetry in CBCT. *Radiat Phys Chem* 167 (2020)
7. Dukov N, Bliznakova K, Okkalidis N, Teneva T, Encheva E, Bliznakov Z (2022) Thermo-plastic 3D printing technology using a single filament for producing realistic patient-derived breast models. *Phys Med Biol* 67(4)
8. Mei K et al (2022) Three-dimensional printing of patient-specific lung phantoms for CT imaging: emulating lung tissue with accurate attenuation profiles and textures. *Med Phys* 49(2):825–835
9. Okkalidis N, Bliznakova K, Kolev N (2022) A filament 3D printing approach for CT-compatible bone tissues replication. *Physica Med* 102:96–102
10. Talalwa L, Natour G, Bauer A, Drzezga A, Beer S (2020) Radiological characteristics of a new experimental rubber elastomeric polymer used in three-dimensional printing with different infill densities and patterns. *J Phys Commun* 4(12) (2020)
11. Savi M, Potiens MDP, Cechinel C, Silveira LCD, Soares FAN (2017) Relationship between infill patterns in 3D printing and Hounsfield Unit. Goiania 30 years later (International Joint Conference RADIO 2017)
12. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9(7):671–675
13. Ma X, Buschmann M, Unger E, Homolka P (2021) Classification of X-ray attenuation properties of additive manufacturing and 3D printing materials using computed tomography from 70 to 140 kVp. *Front Bioeng Biotechnol* 9:763960
14. Kozee M et al (2023) Methodology for computed tomography characterization of commercially available 3D printing materials for use in radiology/radiation oncology. *J Appl Clin Med Phys* 24
15. Ma X, Figl M, Unger E, Buschmann M, Homolka P (2022) X-ray attenuation of bone, soft and adipose tissue in CT from 70 to 140 kV and comparison with 3D printable additive manufacturing materials. *Sci Rep* 12:14580



An Approach to Create a Computational Model of a Trabecular Bone

Kristina Bliznakova^(✉) , Dimitar Trizlov , Nikolay Dukov , Stoyan Ivanov ,
and Zhivko Bliznakov 

Medical University of Varna, Varna, Bulgaria
kristina.bliznakova@mu-varna.bg

Abstract. Spheres are simple 3D geometrical primitives, which when arranged in specific patterns, can form structures exhibiting fractal characteristics. Consequently, spheres are used to construct breast phantoms for x-ray imaging. This study aims to apply this approach of creating complex breast phantoms to bone applications. For this purpose, we utilized realistic tibia outlines based on Computed Tomography patient data and a 3D texture derived from spherical geometrical primitives. Then, these were combined to generate a trabecular bone model. Three bone models were created, and x-ray planar images of these models were produced using a dedicated x-ray imaging simulator. The models were evaluated by orthopedic physicians, who confirmed the realistic appearance of the trabecular structures. The bone model is intended for use in osteoporosis research studies and 3D printing techniques.

Keywords: spherical geometrical primitives · bone model · x-ray images · modelling and simulation

1 Introduction

Spheres are simple 3D geometrical primitives, which when arranged in specific patterns, are able to form structures that exhibit fractal characteristics (Leys, 2005). This fact can be exploited by biomedical researchers to construct realistic anatomical models used in experimental research work. For instance, Gun *et al.* (Gang et al., 2010) constructed a physical phantom based on the properties of fractal self-similarity. The phantom was composed of equal volumes of differently sized spheres, as experimental work demonstrated phantom's power-law spectral density with an exponent of approximately 3. This fact was used by a research group in KU Leuven to develop a structured physical breast phantom for studies related to digital mammography and breast tomosynthesis (Cockmartin et al., 2012; Cockmartin et al., 2013; Cockmartin et al., 2017). The group has built a phantom composed of an acrylic semi-cylinder container of thickness 58 mm and diameter of 200 mm. In addition, equal volumes of acrylic spheres of six different diameters: 15.88 mm, 12.70 mm, 9.52 mm, 6.35 mm, 3.18 mm and 1.58 mm were added in the container. The space between the spheres in the phantom was filled with water in order to represent the x-ray properties of the breast tissue. Investigations suggested that

in order to improve this phantom for x-ray 2D and 3D imaging studies and better match the patients' power spectrum data, the phantom's material contrast should be increased, for example, by introducing paraffin oil (Salomon et al., 2020; Marinov et al., 2021).

To optimize the selection of materials for the spheres and the interstitial fill within the breast phantom, a computational model was developed (Bliznakova, 2017) and successfully used in simulation studies (Baneva et al., 2017). The computational phantom, constructed from tens of thousands of spheres, typically around 40 000, however encounters computational challenges, especially when used with ray tracing techniques to generate synthetic x-ray images. To resolve this issue, a GPU-accelerated procedure was developed and validated to convert the sphere-based computational phantom into a voxel-based model (Bliznakova and Bliznakov, 2024). This transformation enables fast generation of various textures, significantly expanding the applications of sphere-based textures in research studies.

The improved speed of fractal model generation facilitates the development of intricate models mimicking real human tissues. Research has demonstrated that trabecular bone exhibits a fractal dimension, making this method highly adaptable for creating accurate trabecular bone texture models (Chung and Chu, 1994). Our hypothesis is that the sphere-based approach employed in our current research can effectively model the texture of trabecular bone. Therefore, the aim of this study is to adapt our approach for generating realistic breast structures to accurately replicate the trabecular architecture of bone in a 3D bone phantom. The bone model is to be exploited in osteoporosis research studies and 3D printing techniques.

2 Materials and Methods

The main steps for the creation of the computational bone model are shown in Fig. 1. The first step involves creating a suitable 3D texture using geometrical 3D primitives, which are then inserted into the outlined tibia border volume. We selected tibial bone, because it consists of both trabecular and cortical bone. The second step focuses on converting these primitives into a voxel-based representation and obtaining the boundaries of the tibial bone. Next, the generated 3D texture is mapped to the trabecular bone and x-ray images are generated of the created model. Finally, orthopedic physicians evaluated the realism of the reconstructed bone and assessed its potential applications.

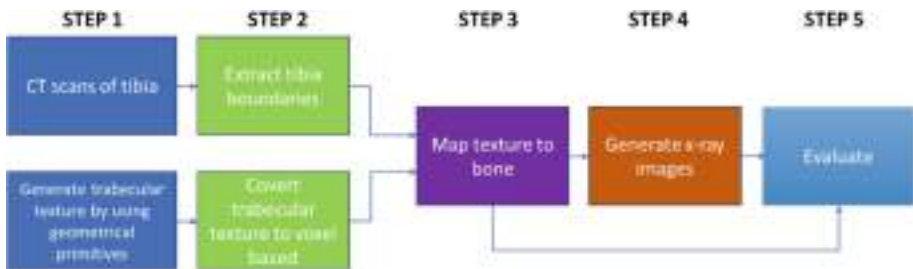


Fig. 1. Main steps in constructing the computational tibial bone model.

2.1 Bone External Shape

The work starts with Computed Tomography (CT) images of tibial bone. The CT images are converted to binary images by applying simple thresholding method in order to segment only the boundaries of the tibia. An example is shown in Fig. 2 for one CT slice. Then, the segmented bone is labeled. The structures of interest are (a) the contours of the segmented bone (the compact bone) and (b) the internal part of the bone, which is the trabecular bone. For this purpose, *imbinarize* and *bwlabel* commands of Matlab 2022a are used. The images in Fig. 2b and Fig. 2c are saved in PNG file format, as the images in Fig. 2c are stacked in a volume, which is the volume used for mapping with the 3D texture, representing the trabecular part of the tibial bone. This step is implemented entirely in Matlab 2022a.

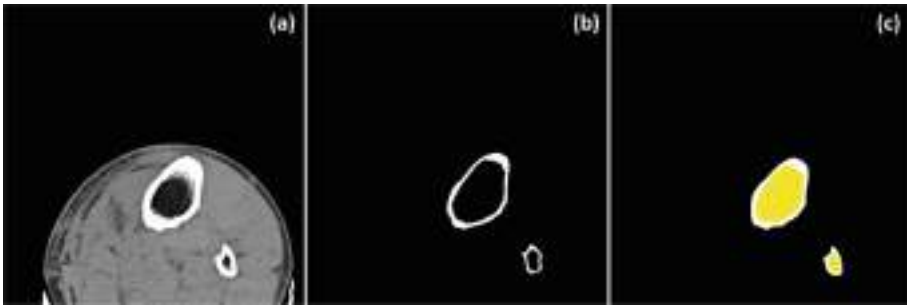


Fig. 2. Creation of tibial bone shape starting from (a) original CT slice of the tibial bone, (b) binary image after thresholding, and (c) labeled trabecular part of the bone.

2.2 Generation of 3D Texture

To generate 3D texture, two processes are applied: (a) generation of a 3D container with 3D spheres of different size, and (b) transforming the generated 3D geometrical primitives to voxel-based volume.

The first process is the generation of a cubic volume with size of 30 mm in each direction. A dedicated software application was created to randomly sample the positions of spherical objects within the cubic volume. The initial diameter, d_{ini} , of the spheres is user-defined and based on this value, each time a new sphere diameter, d_{sphere} , is produced as $rand\_value * d_{ini}$, where *rand_value* is a random value uniformly distributed in the range [0.5 to 1.0]. The algorithm ensures that the place of each sphere is sampled in such a way that it does not intersect with any other spheres. The application is developed in Rad Studio 11 using C++ and input parameters include the radius and color of the spheres, the size and the color of the container. The generated object can be open with the *XRAYImagingSimulator* application (Bliznakova et al., 2010) and can be used in x-ray imaging studies. An example of generated volume of spherical objects with radiuses between 0.25 mm to 0.50 mm is shown in Fig. 3a.

The second step is related to the transformation of the model with the geometrical primitives to voxel-based volume. This has been accomplished via GPU programming.

The programming code has been run onto NVIDIA Quadro RTX 4000, possessing 2304 CUDA cores, which allows for massive parallel execution of tasks. Specifically, each voxel in the 3D grid is treated as an independent task. This means each CUDA core can process a different voxel, in parallel. Further on, in CUDA, threads are grouped into blocks, while blocks are arranged into a grid. Each thread processes one voxel, and the threads are distributed across blocks and grids. The block dimension is $8 \times 8 \times 8$ giving a block size of 512 threads. The converted geometrical matrix is shown in Fig. 3b. The voxel size used in this work is $0.1 \text{ mm} \times 0.1 \text{ mm} \times 0.1 \text{ mm}$.

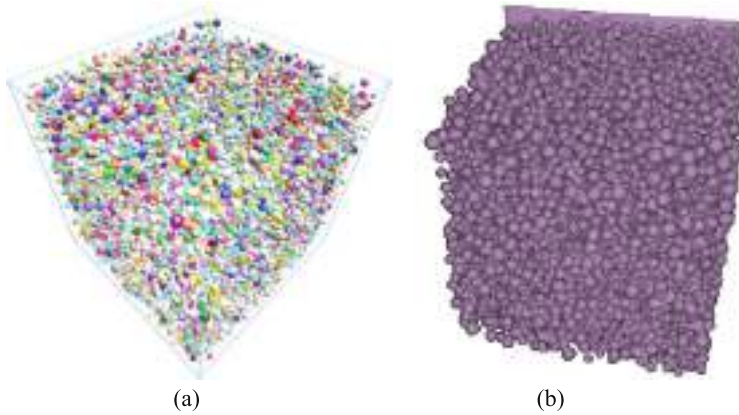


Fig. 3. Converting geometrical texture, from (a) spherical primitives, to (b) voxel based form.

2.3 Complete Bone Phantom and Generation of X-ray Images of the Models

The complete bone phantom is built by mapping the texture, created via GPU programming within the outlined bone volume. The external shape is assigned the properties of the compact bone, while the internal media was assigned trabecular bone properties (Lin and Kang, 2021).

X-ray images of the generated bone phantoms are produced with the *XRAYImagingSimulator* (Bliznakova et al., 2010), by using analytical approach in forming the x-ray projection image. The simulated x-ray imaging setup is demonstrated in Fig. 4, where SSD and SDD are the distances from the x-ray source to the table, where the bone model is placed, and to the detector, respectively. These parameters were set to 600 mm and 800 mm respectively. The size of the detector is set to 2500×2500 pixels; the size of the image pixel is $50 \mu\text{m}$. It was assumed to be composed of a material that fully absorbs incoming photons. The energy of the photons was set to 20 keV. The pixel values correspond to the linear integral, computed using ray tracing techniques.

2.4 Evaluation

A subjective evaluation of the created 3D computer models was performed by two orthopedic physicians. The models were presented with snapshots of each of the four

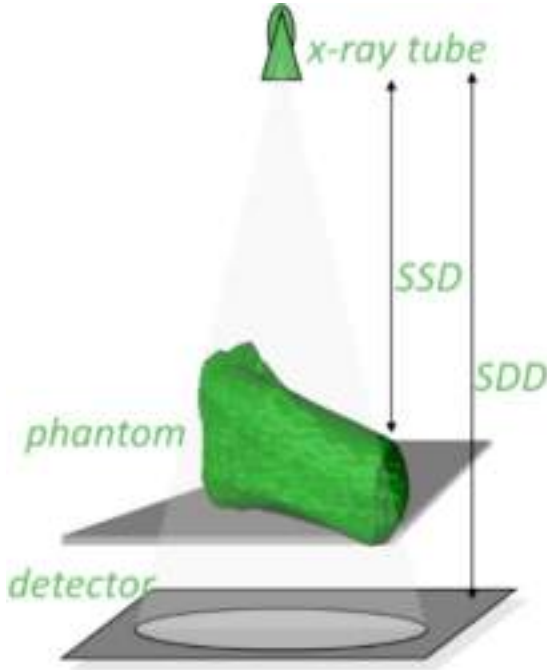


Fig. 4. Experimental setup of the simulation study performed on the 3D bone model.

models from different angles, along with the planar images from the x-ray imaging simulations. Additionally, they were given four identical questions for each model, as shown in Table 1.

This approach was chosen because experts needed to visually assess the obtained trabecular bone pattern and provide feedback to improve the methodology, before quantitative analysis. This approach was also used in our previous studies (Dukov et al, 2021), where initially the realism of the physical phantoms was subjectively assessed.

In addition, the fractal dimension (FD) of the simulated projection images was calculated by using an in-house developed software (Marinov et al, 2021). FD was assessed for a region of interest within the x-ray images of the modelled bone, as shown in Fig. 5.

3 Results

Four tibial bone models were created, by using the same bone shape and different settings for the texture generation, summarized in Table 2. The size of the bone matrix was $189 \times 63 \times 189$ voxels, with voxel size $0.5 \text{ mm} \times 0.8 \text{ mm} \times 0.5 \text{ mm}$.

The 3D computer bone models, generated using the parameters in Table 2, along with their corresponding simulated x-ray projections, are presented in Fig. 6, Fig. 7, Fig. 8, and Fig. 9. In these figures, (a) displays a 3D computer model visualized as a volume in *ImageJ*, (b) shows an internal bone view, (c) presents a slice from the computational bone model, and (d) depicts the generated x-ray projection.

Table 1. Evaluation questionnaire for the created 3D models and the simulated planar x-ray images.

Question	Possible answers
Does the 3D model appear realistic?	A) Yes, it closely resembles real bone structure B) Somewhat, but it lacks finer details C) No, it does not look like a real bone
Which real bone does the model resemble the most?	A) Tibia B) Femur C) Skull D) Other
In your opinion, do the simulated x-ray images look realistic?	A) Yes, they closely resemble real x-ray or CT scans B) Somewhat, but they need better material properties C) No, they look artificial and unrealistic
How do you see this model being applied in your practical work?	A) Medical imaging training and simulation B) Surgical planning for orthopedic procedures C) Biomedical engineering and implant testing D) Bone disease research (e.g., osteoporosis simulations) E) Other

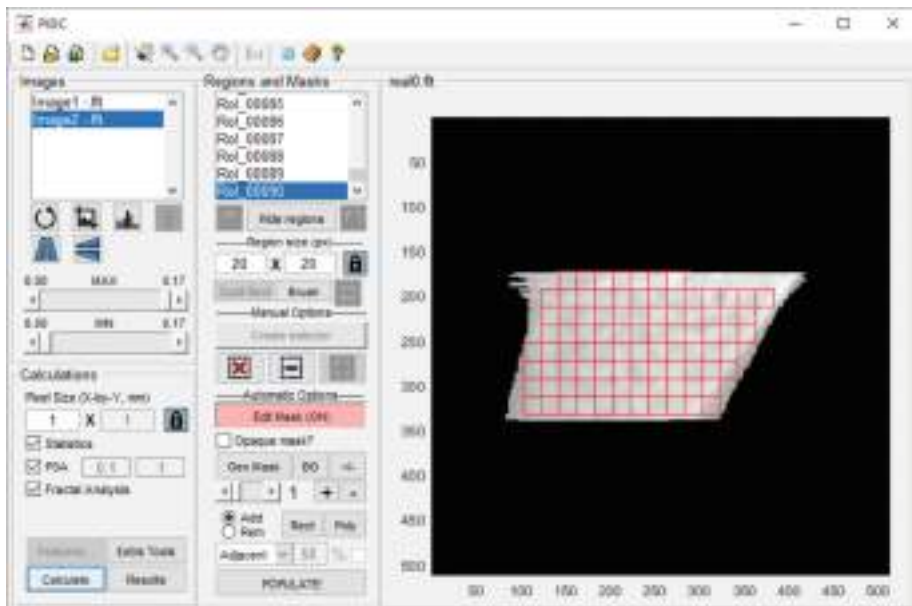
**Fig. 5.** Extracting fractal dimension using a dedicated software tool (Marinov et al., 2021).

Table 2. Parameters, used in the generation of computational bone models.

Model	Number of spheres	Radiuses
Bone 1	46624	(0 - 0.5]
Bone 2	5806	(0.5 - 1]
Bone 3	22391	(0.3 - 0.6]
Bone 4	13030	(0.3 - 0.5]

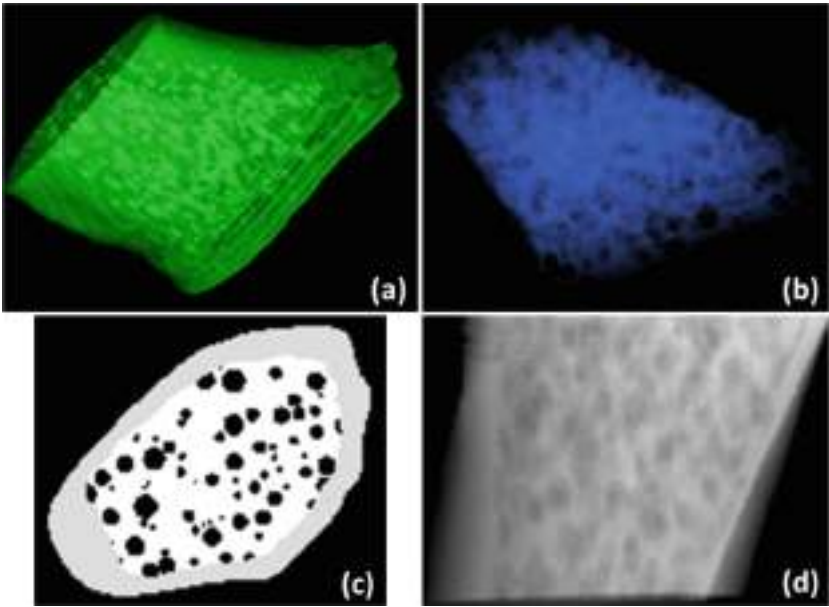


Fig. 6. Model Bone 1: (a) created 3D computational model; (b) interior view of the 3D computational model; (c) slice from the computational model; (d) resulting simulated x-ray planar image of the model.

Spherical structures are observed within the tibial bone across all texture settings. The increased porosity of the simulated trabecular bone in cases 1, 3, 4 is visible both, on the 3D model images as well as on the x-ray generated images. The FD values were ranging from 2.0136 to 2.0144 for all bone models, which are lower than those reported by Inoue and Ogawa (1996), who obtained FD values of 2.48 to 2.86 for trabecular bone structures in the femoral head. The FD is a measure of the complexity and roughness of a structure; therefore, we expect when the texture generation is improved to produce more complex and more irregular structures to result in higher fractal dimension.

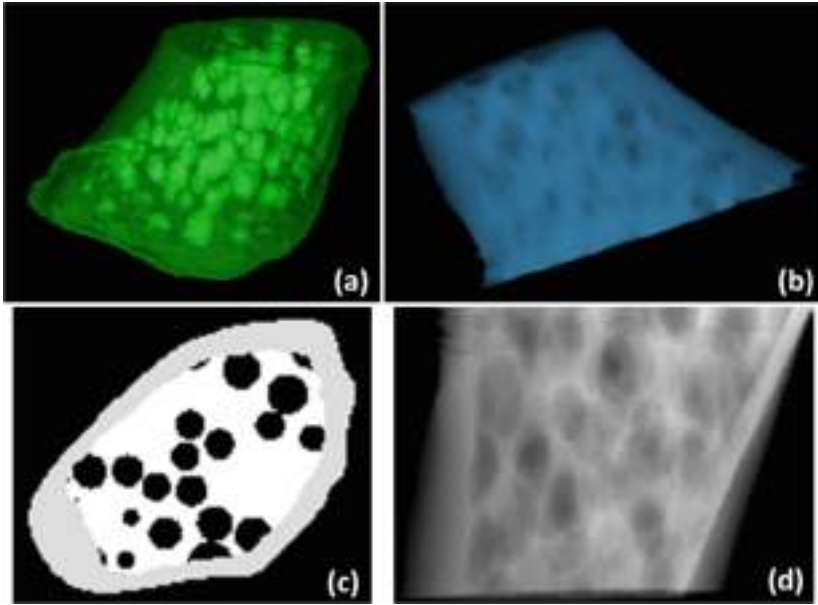


Fig. 7. Model Bone 2: (a) created 3D computational model; (b) interior view of the 3D computational model; (c) slice from the computational model; (d) resulting simulated x-ray planar image of the model.

The two orthopedic physicians evaluated the images and concluded that they resemble realistic bone segments composed of both compact and trabecular bone. They found them highly suitable for research in various medical fields, particularly in studies on osteoporosis, tumor development, and preoperative planning. Their opinion is that these models may represent both, the tibia and the femur, and may be used as a tool for understanding bone pathology and optimizing surgical interventions.

Future work will focus on enhancing texture representation by developing more complex representations of geometric primitives, instead of spheres and conducting an objective evaluation of both, 3D models and x-ray images. Experimental validation is planned, including 3D printing of the computational bone models and their CT scans. For this purpose, we will use reconstructions from bone micro-CT scan. A sample has been already sent to a micro-CT facility. Features, such as skewness, kurtosis, and spectral analysis, will be extracted from CT slices of the scanned real and computational bones via an in-house developed software (Marinov et al, 2021). The features will be then assessed. The research will primarily focus on studies related to osteoporosis.

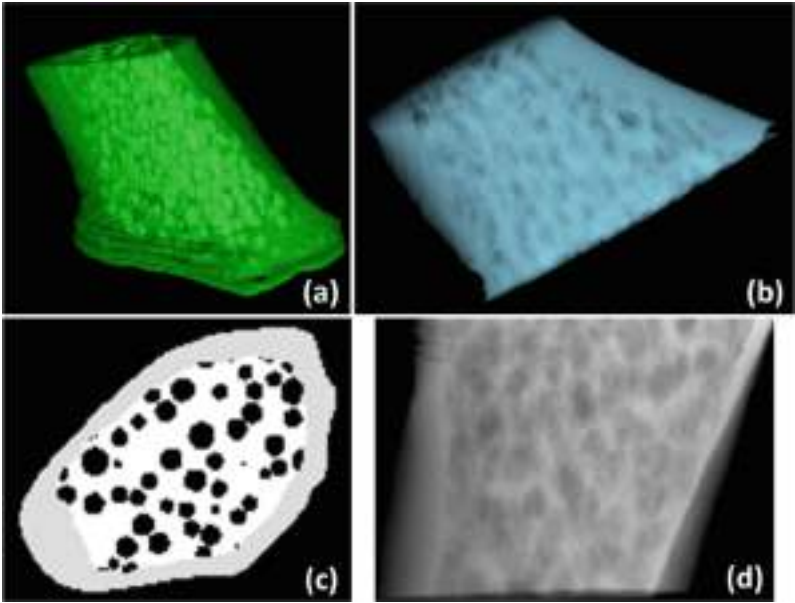


Fig. 8. Model Bone 3: (a) created 3D computational model; (b) interior view of the 3D computational model; (c) slice from the computational model; (d) resulting simulated x-ray planar image of the model.

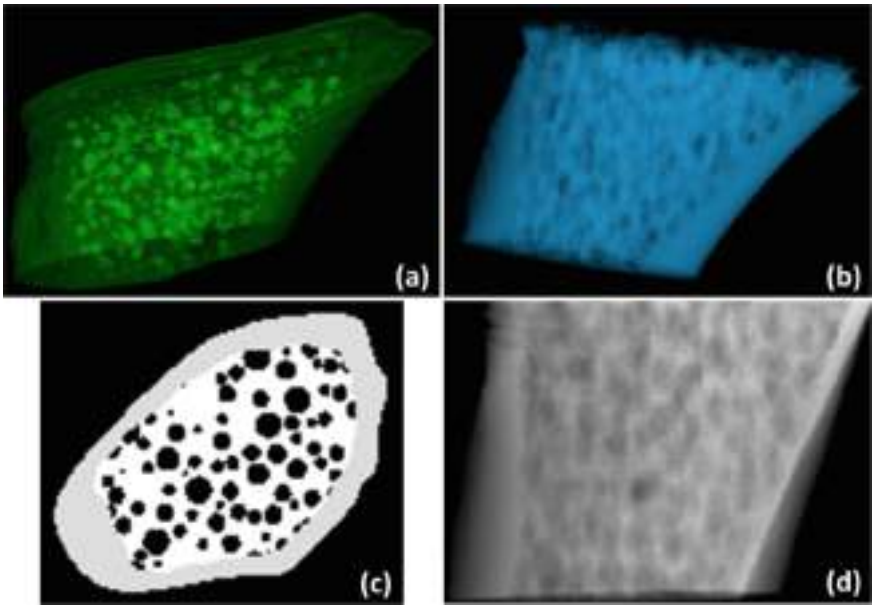


Fig. 9. Model Bone 4: (a) created 3D computational model; (b) interior view of the 3D computational model; (c) slice from the computational model; (d) resulting simulated x-ray planar image of the model.

4 Conclusions

This study tested the hypothesis that the approach used to create realistic 3D textures for breast models, based on spherical geometrical primitives, could be applied to 3D bone models. The hypothesis was successfully demonstrated and the resulting bone models were positively evaluated by our orthopedic physicians. However, the measured fractal dimensions were lower than those reported in the literature, indicating that further work is needed to refine the model in this regard, as well as to perform experimental validation. The bone models are intended for use in osteoporosis research studies and 3D printing techniques.

Acknowledgement. This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0009-C02. The work is performed by the ELPIDA 3.1.5 research group, which is part of the MUVE-TEAM project No BG-RRP-2.004-0009-C02.

Conflict of Interest. The authors declare that they have no conflict of interest.

References

- Baneva Y et al (2017) Evaluation of a breast software model for 2D and 3D x-ray imaging studies of the breast. *Phys Med* 41:78–86. <https://doi.org/10.1016/j.ejmp.2017.04.024>
- Bliznakova K (2017) Development of breast software phantom dedicated for research and educational purposes. *RAD Assoc J* 2:14–19. <https://doi.org/10.21175/RadJ.2017.01.004>
- Bliznakova K, Bliznakov Z (2024) A novel breast phantom created using GPU-based voxelization for x-ray breast imaging research. In: E-Health and Bioengineering Conference (EHB), IASI, Romania, pp 1–4. <https://doi.org/10.1109/EHB64556.2024.10805680>
- Bliznakova K, Speller R, Horrocks J, Liaparinos P, Kolitsi Z, Pallikarakis N (2010) Experimental validation of a radiographic simulation code using breast phantom for x-ray imaging. *Comput Biol Med* 40:208–214. <https://doi.org/10.1016/j.compbiomed.2009.11.017>
- Chung, H., Chu, C., Underweiser, M., Wehrli, F.: On the fractal nature of trabecular structure. *Med. Phys.* **21**, 1535–1540 (1994). <https://doi.org/10.1118/1.597263>
- Cockmartin L, Marshall N, Bosmans H (2012) Design and evaluation of a phantom with structured background for digital mammography and breast tomosynthesis. In: Maidment ADA, Bakic PR, Gavenonis S (eds) *Breast Imaging. IWDM 2012. Lecture Notes in Computer Science*, vol 7361. Springer, Heidelberg, pp 642–649. https://doi.org/10.1007/978-3-642-31271-7_83
- Cockmartin L, Bosmans H, Marshall NW (2013) Comparative power law analysis of structured breast phantom and patient images in digital mammography and breast tomosynthesis. *Med Phys* 40:081920. <https://doi.org/10.1118/1.4816309>
- Cockmartin L et al (2017) Design and application of a structured phantom for detection performance comparison between breast tomosynthesis and digital mammography. *Phys Med Biol* 62:758–780. <https://doi.org/10.1088/1361-6560/aa5407>
- Dukov N et al (2021) Experimental evaluation of physical breast phantoms for 2D and 3D breast x-ray imaging techniques. In: Jarm T, Cvetkoska A, Mahnič-Kalamiza S, Miklavcic D (eds) *8th European Medical and Biological Engineering Conference. EMBEC 2020. IFMBE Proceedings*, vol 80. Springer, Cham, pp 544–552. https://doi.org/10.1007/978-3-030-64610-3_62

- Gang GJ, Tward DJ, Lee J, Siewerdsen JH (2010) Anatomical background and generalized detectability in tomosynthesis and cone-beam CT. *Med Phys* 37:1948–1965. <https://doi.org/10.1118/1.3352586>
- Inoue E, Ogawa K (1995) Analysis of trabecular patterns using fractal dimensions. In: *IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, San Francisco, CA, USA, 3, pp 1497–1500. <https://doi.org/10.1109/NSSMIC.1995.500310>
- Leys J (2005) Sphere inversion fractals. *Comput Graph* 29:463–466. <https://doi.org/10.1016/j.cag.2005.03.011>
- Lin CY, Kang JH (2021) Mechanical properties of compact bone defined by the stress-strain curve measured using uniaxial tensile test: a concise review and practical guide. *Materials* 14(15):4224. <https://doi.org/10.3390/ma14154224>
- Marinov S et al (2021) Radiomics software for breast imaging optimization and simulation studies *Phys. Medica* 89:114–128. <https://doi.org/10.1016/j.ejmp.2021.07.014>
- Salomon E et al (2020) Equivalent breast thickness and dose sensitivity of a next iteration 3D structured breast phantom with lesion models. *Proc. SPIE* 11312, *Medical Imaging 2020: Physics of Medical Imaging*, 113120K. <https://doi.org/10.1117/12.2548956>



Explainable Estimation of Blood Volume Pulse Signals from Video Sequences Using a Combination of Deep Learning Models and Signal Processing Methods

Milena Sobotka^(✉), Kamil Kopryk, Muhammad Usman, and Jacek Rumiński

Department of Biomedical Engineering, Gdansk University of Technology, FETI,
80-233 Gdansk, Poland
milena.sobotka@pg.edu.pl
<http://www.pg.edu.pl>

Abstract. This study evaluates the application of Gradient-weighted Class Activation Mapping (Grad-CAM) to identify key image regions for enhancing pulse wave estimation using traditional signal analysis methods. The approach assumes that Grad-CAM masking enables the selection of relevant image areas, improving the Signal to Noise Ratio (SNR). Experiments were conducted on the PURE dataset using the TS-CAN model and rPPG-Toolbox. Grad-CAM maps identified facial regions most influential for model prediction, allowing the exclusion of areas not contributing to accurate estimation. The study also explored different temporal window sizes and their impact on signal quality. Method evaluation included SNR_{raw} , SNR_{max} , SNR_{sum} , and Hjorth descriptors. Results confirmed that Grad-CAM masking enhances rPPG signal quality, enabling more precise heart rate estimation. Statistical analysis validated the significance of the findings, highlighting the potential of interpretable deep learning methods in signal analysis.

Keywords: blood volume pulse · remote photoplethysmography · signal processing · explainable AI

1 Introduction

Technological advancements have enabled the development of non-invasive methods for monitoring physiological parameters, among which remote photoplethysmography (rPPG) plays a crucial role. This technique utilizes subtle changes in skin color, captured in video recordings, to analyze pulsatile blood flow, allowing for the estimation of heart rate (HR) and blood volume pulse (BVP) [1]. These variations result from small fluctuations in blood volume during the cardiac cycle, which influence the amount of light absorbed by blood vessels and, consequently, cause alterations in skin intensity values. Unlike traditional contact-based methods, such as photoplethysmography (PPG), rPPG eliminates the

need for attached sensors, significantly enhancing user comfort and enabling continuous, remote monitoring. As a result, it finds wide application in home-based health monitoring systems. Despite its advantages, rPPG still encounters challenges related to physiological factors influencing BVP signal measurement. One of the key obstacles is the spatial variability of the vascular system and the uneven distribution of signal intensity across different regions of interest (ROI) on the face. Furthermore, the low amplitude of these signals makes them particularly vulnerable to motion artifacts and lighting fluctuations, often causing noise levels to exceed the intensity of the useful signal. In response to these challenges, various filtering algorithms and noise reduction techniques have been developed [2–4]. Among the most commonly used methods are those based on independent component analysis (ICA), which allow for the decomposition of RGB signals into independent components, from which those containing relevant heart rate information are selected. Alternatively, models are employed where RGB channels are represented as a linear combination of BVP components and noise, allowing for their separation based on prior assumptions. Deep learning-based methods [5,6] have achieved significant importance in analyzing video data to extract physiological parameters such as HR and BVP. Once trained on large datasets, these models can effectively predict HR components, thereby improving measurement accuracy. Accurate monitoring of vital parameters requires advanced image processing techniques combined with deep learning methods, which enable improved measurement precision, especially under challenging environmental conditions.

The traditional signal processing based methods used to estimate the pulse wave signal are fully explainable. However, usually the fixed masks for face, forehead or cheeks produce low quality signals. One of the key challenges in applying deep learning to medical contexts is its limited interpretability. Although these models can efficiently extract high-quality signals from input video, they often function as black boxes, raising concerns among medical professionals. In response to these issues, the field of Explainable Artificial Intelligence (XAI) is being developed, focusing on creating interpretable models while maintaining high predictive accuracy. One widely used solution is Grad-CAM [7], which allows for the visualization of image areas with the most significant impact on the model’s decision. Grad-CAM has found applications in various aspects of facial analysis [8,9], including identity recognition, emotional expression analysis, attribute classification, and medical diagnostics [10,11].

This paper aims to combine both techniques by proposing a methodology that uses Grad-CAM to identify key image regions, enhancing pulse wave estimation accuracy through traditional signal analysis. The main assumption is that Grad-CAM-based masking enables the selection of the most relevant image areas, thereby increasing the SNR compared to analyzing entire images. To evaluate this approach, the Temporal Shift Convolutional Attention Network (TS-CAN) [17] was assessed using the rPPG-Toolbox [16] on the Pulse Rate Detection Dataset (PURE) [24]. Grad-CAM heatmaps were then generated to determine which facial regions had the most significant influence on the model’s predic-

tions, allowing for the identification of the most critical areas for rPPG signal extraction, and using these heatmaps, a binary mask was created. Additionally, a windowed approach to mask generation was introduced, incorporating varying numbers of neighboring frames in the computation of Grad-CAM heatmaps, and its impact on heart rate estimation accuracy was analyzed. The effectiveness of the proposed method was assessed using various metrics, including the SNR_{raw} , SNR_{max} , SNR_{sum} , and Hjorth descriptors. Moreover, statistical analyses were conducted to determine the significance of the results and to assess the influence of the applied methods on signal quality.

The rest of the paper is structured as follows. Section 2 provides a review of existing rPPG methodologies, with a particular focus on the impact of lighting conditions, ROI selection, and the application of deep neural networks in heart rate estimation. Section 3 describes the proposed methodology, including the implementation of Grad-CAM-based masking and the applied signal analysis techniques. Section 4 presents the experimental results and statistical analysis, followed by a discussion of the findings in Sect. 5. Finally, Sect. 6 concludes the paper by discussing potential improvements in rPPG signal acquisition.

2 Related Work

In recent years, rPPG has gained recognition as a progressive method for contactless monitoring of vital parameters. Research in this field focuses on identifying factors that affect measurement accuracy and improving signal processing methods. The study [12] analyzed the impact of various lighting conditions, frame rates, and video compression on the effectiveness of heart rate detection using videoplethysmography. It was shown that different ambient light sources significantly influence the quality of pulse estimation in rPPG, and optimal lighting conditions, along with proper camera settings, are crucial for achieving accurate measurements. The findings also emphasize the necessity of minimizing disturbances caused by video compression and further developing rPPG technology to address the challenges of real-world environmental conditions. The study [3] demonstrated the feasibility of measuring pulse rate using a webcam by analyzing color changes in specific facial regions, such as the forehead. The authors proposed focusing on selected areas of the face to enhance measurement accuracy and reduce the impact of disturbances. The study highlighted the effectiveness of the Principal Component Analysis (PCA) method in extracting the pulse signal while ensuring high computational efficiency, making it suitable for real-time applications. It was also noted that selecting smaller ROI helps minimize motion artifacts and provides reliable measurements under controlled lighting conditions. Further research [13] explored the use of normalized and tracked facial regions, such as the cheeks and nose, to improve signal stability and enhance the robustness of rPPG measurements. Recent advancements in rPPG research focus on leveraging deep neural networks to enhance the precision and reliability of pulse signal analysis. The study [14] demonstrated that the application of DNNs enables not only the classification of PPG signals based on

their suitability for heart rate estimation but also the automatic identification of facial regions that provide the most stable measurements. Additionally, signal quality analysis, incorporating parameters such as SNR and Normalized Peak Energy (NPE), further improves classification accuracy and reduces the likelihood of erroneous readings. Unlike traditional methods such as ICA and PCA, which may produce false-positive results in the presence of noise, deep learning models demonstrate greater adaptability to varying lighting conditions while effectively minimizing motion artifacts. Furthermore, the use of DNN regression models allows for direct heart rate estimation without requiring complex filtering techniques, making this approach particularly promising for telemedicine applications and intelligent health monitoring systems.

One of the key aspects of modern rPPG methods is the utilization of attention mechanisms and deep learning architectures, as demonstrated in the study by [15]. The introduced MAR-rPPG method focuses on improving the precision of ROI localization and enhancing motion robustness. The implementation of masked attention enables increased semantic consistency in attention maps across consecutive video frames, while the masking technique prevents the model from overly relying on imprecisely determined ROIs. Additionally, the integration of the Enhanced rPPG Expert Aggregation (EREA) architecture allows for a dynamic analysis of various facial regions and selectively directs the model's attention to the most relevant features of the rPPG signal. The findings indicate that combining attention mechanisms with the elimination of incorrectly designated ROIs significantly enhances rPPG's resistance to motion artifacts, thereby improving the stability and accuracy of heart rate estimation. A comparable approach was adopted in [16], which introduced rPPG-Toolbox, a tool designed to assess and compare the effectiveness of different rPPG methods. The authors focused on deep learning models that employ convolutional neural networks and attention mechanisms for signal analysis.

Among the discussed architectures, TS-CAN was highlighted for its use of temporal shifts in signal analysis [17], PhysNet for its reliance on three-dimensional convolutional networks [18], DeepPhys for integrating attention mechanisms to select key features [19], and PhysFormer, which incorporates transformers for spatiotemporal analysis [20]. The application of these models enables more effective suppression of motion artifacts and improves heart rate estimation accuracy. Notably, the authors emphasize that the standardization of training processes and the analysis of signal quality impact on predictive performance contribute to the optimization of deep learning methods in rPPG. A significant feature of the rPPG-Toolbox is also its ability to integrate model interpretability mechanisms, such as Grad-CAM, which facilitates the identification of facial regions crucial for heart rate prediction [7].

An essential aspect of rPPG analysis is assessing signal quality, as it directly impacts the accuracy of heart rate estimation. The quality measure of rPPG signals is often represented by the value of the SNR. In many papers, the SNR quality measure for rPPG is defined as [21]:

$$SNR = 10 \log_{10} \frac{\sum_{f=30}^{240} (U_w(f) S_f)^2}{\sum_{f=30}^{240} ((1 - U_w(f)) S_f)^2}, \quad (1)$$

where f denotes the frequency in bpm, $S(f)$ represents the spectrum of the extracted signal, $U_w(f)$ denotes the binary template window with the window size represented by w :

$$U_w(f) = \begin{cases} 1 & f_{PR} - \frac{w}{2} \leq f \leq f_{PR} + \frac{w}{2} \\ 1 & 2f_{PR} - \frac{w}{2} \leq f \leq 2f_{PR} + \frac{w}{2} \\ 0 & otherwise \end{cases} \quad (2)$$

and f_{PR} represents the true pulse rate value. The values $f = 30$ and $f = 240$ represent the lower and upper limits for possible pulse rates (0.5 Hz, 4 Hz). However, these frequencies are defined differently in some other papers (e.g., [22]). Using such a template window is a good proposition, assuming we know the reference value of the pulse rate. Some methods [23] propose extracting the pulse signal from image sequences using the SNR measure as an objective function during the training of the deep neural networks. This is a very interesting approach. However, there is uncertainty about whether false positive signals could be extracted to fulfill that criterion. The Hjorth descriptors were originally proposed for on-line analysis of EEG signals [25]. They are often used to represent the signal dynamics and purity in reference to the sinus function. These parameters will be used also in this study.

3 Method

In the following section, we provide details about the PURE dataset, including the experimental conditions, categorization of motion scenarios performed by participants, and the data selection process. We also describe the methodology for generating Grad-CAM masks and calculating signal quality metrics, such as SNR, Hjorth descriptors, which were used in the analysis.

3.1 Dataset

For the purpose of this study, data from the PURE dataset were utilized [24], including recordings of participants performing controlled head movements in front of a camera, while both facial images and reference heart rate measurements were captured. The participants were positioned approximately 1.1 m from the camera under natural daylight conditions. Each participant (01–10) engaged in six distinct scenarios, including: (1) remaining motionless, (2) speaking, horizontal head movements at different speeds – (3) slow and (4) fast translation, as well as looking at designated points around the camera, which induced (5) small and (6) medium head rotations ranging from 20° to 35° . A file naming convention $AA-BB$ was adopted, where AA represented the participant index and BB denoted the scenario number, e.g., 01 – 01. During the recordings, participants'

heart rates ranged from 42 BPM to 148 BPM, with all measurements obtained in a resting state. Due to issues with face detection during data preprocessing for TS-CAN inference and face mask generation, the sample for participant 05 and sample 06 – 02 were removed from the analysis.

3.2 Explainability Analysis of the TS-CAN Model

The TS-CAN model was trained on the PURE dataset for contactless monitoring of physiological parameters, such as heart rate and respiratory rate, through video analysis. A fundamental component, the temporal shift module, efficiently models dynamic signal variations while maintaining low computational costs. This enables real-time operation on devices with limited processing power. Its architecture consists of two primary branches: the appearance branch, which detects subtle skin tone variations indicative of blood flow, and the motion branch, which captures micro-movements related to respiration and cardiac activity. Additionally, the attention module enhances signal to noise separation by focusing on the most relevant facial regions. The use of multi-task learning allows for the simultaneous estimation of heart rate and respiratory rate, optimizing computational efficiency and improving model accuracy.

The training process was conducted using rPPG-Toolbox, an open-source framework dedicated to rPPG signal analysis. The model was trained for 40 epochs with a batch size of 4 and a learning rate of $9 \cdot 10^{-3}$, achieving the best performance in epoch 34. Data from the PURE dataset underwent preprocessing, including DiffNormalization and Standardization. During initial processing, the facial region was detected in the first frame of each video, cropped, and expanded by 50% around the detected area. These regions were then resized to 72×72 px and used in both training and inference to generate Grad-CAM visualizations. In the conducted experiments, two masks were introduced: 1) mask m_1 , which defines a rectangular facial region, and 2) mask m_2 , derived from Grad-CAM heatmaps. First, using the MediaPipe detector and input frames in the RGB domain, employed for Grad-CAM computation, the facial region was identified, and a rectangular binary mask (m_1) was generated. Subsequently, Grad-CAM maps were computed for the Attention Mask 2 layer, normalized for each frame, and converted into binary masks using a 0.5 threshold. To enhance stability and facilitate comparative analysis, a windowed averaging approach was applied, where consecutive Grad-CAM maps were averaged to construct m_2 . The window size was set to 1, 3, 5, 9, 13, 17, 21, 25, 29, and 33. For example, a mask size of 5 indicates that the Grad-CAM mask for a given frame is calculated using the two preceding, the target, and the two subsequent frames. When data from Attention Mask 2 were unavailable, the window size was progressively increased until the missing information was retrieved. The signal was computed based on the mean intensity of the green channel from input frames, multiplied by both masks. This resulted in two signal variants:

- s_1 , where only the facial mask was applied,
- s_2 , where both the Grad-CAM mask and the facial mask were applied simultaneously.



Fig. 1. Example of a single frame with different applied binary masks

Due to substantial variations in facial region size and the number of unmasked pixels across frames, masked pixels were excluded from signal computation. This prevented their influence on the computed averages, ensuring a more reliable signal comparison (Fig. 1).

3.3 Signal Quality Assessment

The next stage of the analysis involved computing SNR_{raw} , SNR_{sum} and SNR_{max} on the two extracted signals: s_1 and s_2 . As part of the computation of the SNR_{raw} for the entire signal spectrum, a high-pass Butterworth filter was first applied with a cutoff frequency of 0.2 Hz, as lower frequencies were considered as baseline drift. Subsequently, a discrete Fourier transform (FFT) was performed on the filtered signal, resulting in the computed frequency spectrum $X(k)$, which was then analyzed in terms of its absolute value $|X(k)|$. The frequency range of [0.7, 4.0] Hz was considered as the main component of the analyzed pulse signal, while frequencies outside this range were classified as a noise. The signal power, was determined by summing the squared values within the specified frequency range. Similarly, the noise power, was defined as the sum of squared values for frequencies outside the [0.7, 4.0] Hz range. Finally, SNR_{raw} was defined using the (Eq. 1) formula, with the values $f_{min} = 42$ and $f_{max} = 240$ (beats per minute corresponding to 0.7–4 Hz) and $U_{wi}(f)$ (Eq. 3), where $U_{wi}(f)$ denotes the binary template window with the window size represented by w for the particular position of the template window i :

$$U_w(f) = \begin{cases} 1 & f - \frac{w}{2} \leq f \leq f + \frac{w}{2} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

SNR_{sum} (Eq. 4) and SNR_{max} (Eq. 5) measures refer to the maximum peak in the spectrum (describing the “energy” of the dominating frequency) without reference to the unknown, ground-truth pulse rate value. The values of these measures reflect how much some frequency (within the range of e.g., ± 5 bpm, if $w=3$) dominates in the spectrum, even if this fundamental frequency is not related to the pulse wave. This approach uses only information within measured signal in the potentially useful bandwidth, without unknown pulse rate (Eq. 1).

In our case, we assumed a frequency range of 0.7–4.0 Hz and a window size of 15 for the calculations.

$$SNR_{sum} = 10 \log_{10} \sum_{i=1}^{N-w} \frac{\sum_{f=42}^{240} (U_{wi}(f) S_f^2)^2}{\sum_{f=42}^{240} ((1 - U_{wi}(f)) S_f^2)^2}, \quad (4)$$

$$SNR_{max} = 10 \log_{10} (\max_{i=1}^{N-w} \frac{\sum_{f=42}^{240} (U_{wi}(f) S_f^2)^2}{\sum_{f=42}^{240} ((1 - U_{wi}(f)) S_f^2)^2}) \quad (5)$$

Activity, mobility, and complexity Hjorth descriptors (Eq. 6) were adapted for the input signal y as:

$$\begin{aligned} activity &= w_0 = \text{var}(y), & mobility &= \sqrt{\frac{w_2}{w_0}} = \sqrt{\frac{\text{var}(\dot{y})}{\text{var}(y)}}, \\ complexity &= \sqrt{\frac{w_4}{w_0}} = \sqrt{\frac{\text{var}(\ddot{y})}{\text{var}(y)}}, \end{aligned} \quad (6)$$

where $\dot{y}$ and $\ddot{y}$ are the first and second derivative of the signal y . In the case of a discrete-time signal, activity is defined by the signal's variance, representing its total energy, while mobility describes the dominant frequency. The complexity (different than proposed by Hjorth) measures the high dynamics of the signal. For the analyzed domain of the signal these descriptors have lower values when variance of noise is low.

4 Results

The impact of different mask types and window sizes on signal quality was evaluated. The analysis was focused on computed SNR metrics and Hjorth descriptors, providing insights into performance differences between masks m_1 and m_2 . The results were summarized in tables and figures, with Fig. 2 illustrating frames where the proposed Grad-CAM masks were applied.



Fig. 2. Sequence of frames with extracted m_2 masks

4.1 Signal to Noise Ratio Analysis

The calculated SNR metrics for different mask types were presented in Table 1, which reported values for SNR_{raw} , SNR_{max} , and SNR_{sum} . The table included the differences between these metrics for m_1 and m_2 .

Table 1. Comparison of SNR_{raw} , SNR_{max} , and SNR_{sum} metrics

	m_1			m_2			$m_2 - m_1$	
	SNR_{raw}	SNR_{max}	SNR_{sum}	SNR_{raw}	SNR_{max}	SNR_{sum}	$\Delta\text{SNR}_{\text{raw}}$	$\Delta\text{SNR}_{\text{max}}$
1	0.1118	-0.9131	9.368	4.221	8.726	17.57	4.109	9.639
2	-0.6648	-8.272	4.549	-0.3398	-4.19	8.115	0.325	4.082
3	-2.592	-6.131	5.625	-0.1757	-2.208	8.505	2.416	3.923
4	-2.107	-5.653	5.975	0.1769	-6.374	5.749	2.284	-0.721
5	-5.082	-4.627	6.259	-1.886	1.646	11.91	3.196	6.273
6	-4.781	-5.7	5.963	-3.247	1.627	11.39	1.534	7.327

To further analyze the impact of mask types, Table 2 provided comparative results across all categories, incorporating Hjorth descriptors and their differences for both masks.

Table 2. Comparison of Hjorth descriptors: activity, mobility, and complexity

	m_1			m_2			$m_2 - m_1$		
	activity	mobility	complexity	activity	mobility	complexity	$\Delta\text{activity}$	$\Delta\text{mobility}$	$\Delta\text{complexity}$
1	0.0001217	0.8115	1.334	4.277e-05	0.4664	0.6836	-7.4e-05	-0.3451	-0.6504
2	0.000247	0.612	0.9224	0.0001148	0.44	0.6078	-0.0001322	-0.172	-0.3146
3	0.001994	0.5858	0.9133	0.000993	0.5118	0.7366	-0.001001	-0.074	-0.1767
4	0.0004462	0.5629	0.8821	0.000166	0.5985	0.8774	-0.0002802	0.0356	-0.0047
5	0.0003806	0.4999	0.8056	8.876e-05	0.4134	0.6102	-0.0002918	-0.0865	-0.1954
6	0.0004796	0.4977	0.7921	0.0001968	0.3765	0.5532	-0.0002828	-0.1212	-0.2389

Table 3 summarized the statistical tests used to compare signals derived from m_1 and m_2 , including the Mann-Whitney test (MW p-value and Rank Biserial Correlation), as well as the Shapiro-Wilk test results for groups $G1$ and $G2$. Additionally, the t-test, applicable under normal distribution assumptions, was included.

Table 3. Comparison of statistical metrics for SNR_{raw} and SNR_{max}

	SNR_{raw}					SNR_{max}				
	MW p-val	MW RBC	SW G_1	SW G_2	TT p-val	MW p-val	MW RBC	SW G_1	SW G_2	TT p-val
1	0.0047	−0.8025	0.7928	0.9631	0.0012	0.0081	−0.7531	0.0293	0.2293	0.0044
2	0.8785	−0.0625	0.4058	0.2693	0.6107	0.2345	−0.3750	0.0453	0.3495	0.1353
3	0.0273	−0.6296	0.7643	0.2208	0.0274	0.0774	−0.5062	0.0282	0.5354	0.1322
4	0.0217	−0.6543	0.8759	0.7173	0.0164	0.8598	0.0617	0.3016	0.7395	0.7526
5	0.0423	−0.5802	0.2820	0.1148	0.0232	0.0171	−0.6790	0.1524	0.2598	0.0140
6	0.2893	−0.3086	0.2879	0.2378	0.2007	0.0020	−0.8765	0.1512	0.4680	0.0005

4.2 Window Size Comparison

The effect of window size on SNR_{raw} and SNR_{max} in the computation of the Grad-CAM binary mask was examined. Figures 3 and 4 displayed SNR_{raw} for m_2 alongside baseline values for m_1 . Figures 5 and 6 illustrated the differences in SNR_{raw} and SNR_{max} , highlighting the impact of varying the window size on signal quality.

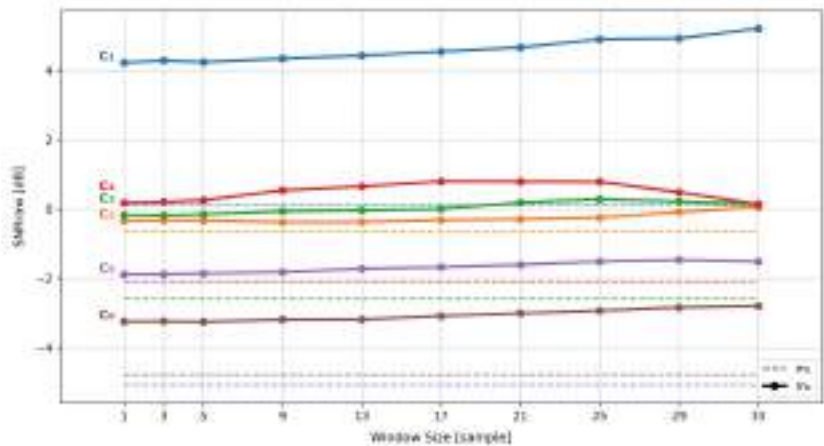


Fig. 3. Comparison of SNR_{raw} for two masks across different window sizes

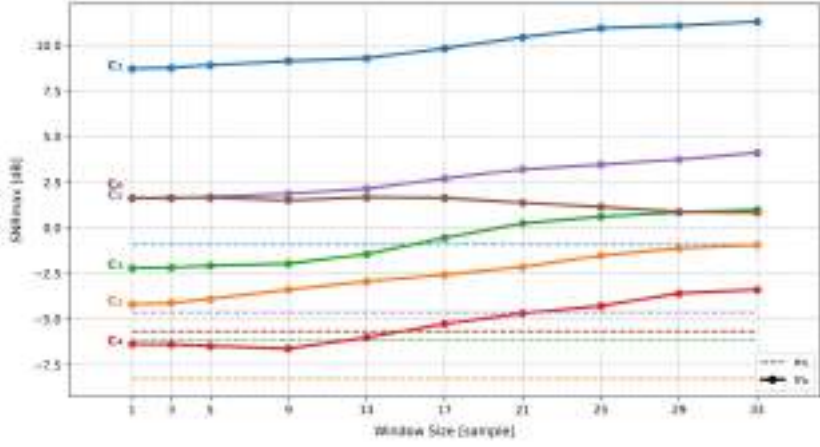


Fig. 4. Comparison of $\text{SNR}_{\max}$ for two masks across different window sizes

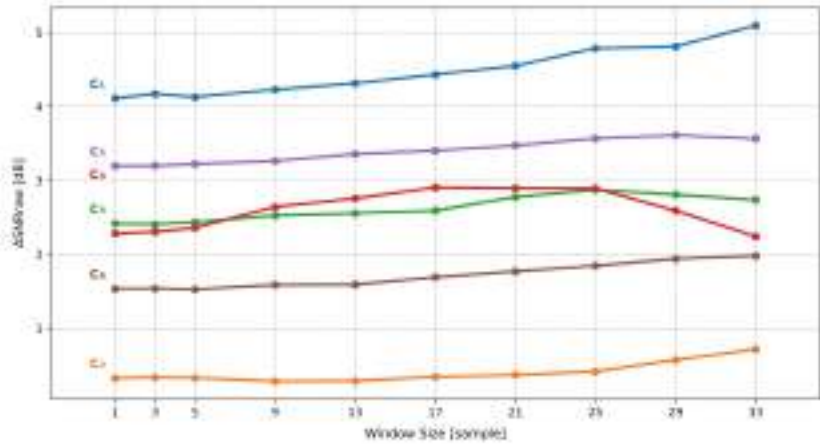


Fig. 5. Comparison of $\text{SNR}_{\text{raw}}(m_2) - \text{SNR}_{\text{raw}}(m_1)$ across different window sizes

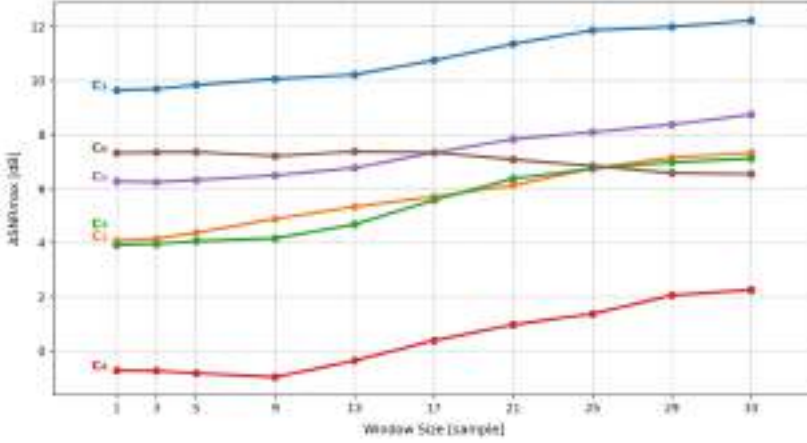


Fig. 6. Comparison of $\text{SNR}_{\max}(m_2) - \text{SNR}_{\max}(m_1)$ across different window sizes

5 Discussion

The largest improvements in SNR_{raw} with Grad-CAM-based masking were observed in category 1 (no movement), where the SNR increased by +4.109, and in category 5 (small head rotations), with an increase of +3.196. Notable increases occurred in categories 3 (slow head movement, +2.416), 4 (rapid head movement, +2.284), and 6 (large head rotations, +1.534). Category 2 (speaking activity) showed minimal improvement (+0.325), suggesting that speech artifacts persist despite masking. $\text{SNR}_{\max}$ improved notably, with the largest increase in categories 1 (+9.639), 6 (+7.327) and 5 (+6.273). Moderate improvements were seen in categories 2 (+4.082) and 3 (+3.923), whereas category 4 showed a slight decline (−0.721), indicating potential distortions under fast motion using Grad-CAM masking. In our Hjorth descriptors implementation, all differences between masks m_2 and m_1 were negative, indicating reduced noise in the obtained signal, except for a slight mobility increase in one instance of category 4. Considering the temporal window size of the attention mask in relation to the increase in SNR_{raw} and $\text{SNR}_{\max}$, the findings indicate that SNR_{raw} exhibits the greatest improvement in category 1 (+0.9840), while in categories 2, 3, and 5, the differences remain minimal, indicating that the expansion of the window led to only a slight improvement in signal enhancement. In contrast, $\text{SNR}_{\max}$ shows a significantly greater improvement, particularly in categories 1, 2, 3, 4 and 5, where differences exceed +2, suggesting that larger window sizes substantially improved signal quality in the 0.7–4.0 Hz band.

Based on the Mann-Whitney and Shapiro-Wilk tests, significant differences between the analyzed groups were observed in certain categories, with both statistical significance (MW p-value) and effect size (MW RBC) playing a crucial role in result interpretation. For SNR_{raw} , significant differences were found in

categories 1, 3, 4, and 5, as confirmed by the low MW p-values (MW p-val < 0.05) and large effect sizes ($|\text{MW RBC}| \geq 0.5$), confirming substantial differences between groups. Category 6, despite demonstrating a moderate effect size ($0.3 \leq |\text{MW RBC}| < 0.5$), did not reach statistical significance in the Mann-Whitney test, suggesting that the observed differences may be attributed to random data fluctuations rather than systematic variance. Meanwhile, category 2 exhibited neither statistical significance (MW p-val = 0.8785) nor a meaningful effect size ($|\text{MW RBC}| < 0.3$). For SNR_{max} , significant differences were observed in categories 1, 5, and 6, where (MW p-val < 0.05) and effect sizes were large ($|\text{MW RBC}| \geq 0.5$), indicating a strong distinction between the s_1 and s_2 signals. Category 2, despite reaching a moderate effect size ($0.3 \leq |\text{MW RBC}| < 0.5$), did not demonstrate statistical significance, while category 4 showed both a small effect size and no statistical significance. Category 3 exhibited a large effect size ($|\text{MW RBC}| = 0.5062$), but did not reach statistical significance.

This preliminary study has many limitations. First, it uses only one public dataset. Many other datasets can be analyzed in the future. Second, only one deep-learning model was used to extract Grad-CAM arrays and calculate masks used in this study. The TS-CAN model is very effective; however, more models have been proposed. Third, in our study, we considered only the signal from the green channel, whereas many traditional techniques, such as ICA and PCA, can significantly enhance signal quality by utilizing all three available channels. Nevertheless, this study presents a case showing how we can extract regions of a face that contribute most to the final, estimated BVP signal. It can be extended to propose a fully automated method to extract BVP-related face regions instead of fixed ROIs like the face area, forehead, cheeks, etc.

6 Conclusion

The performed analyses indicated that the application of Grad-CAM-based masking leads to a significant improvement in the quality of the rPPG signal. In addition to enhancing signal quality, the application of a binary mask derived from Grad-CAM, generated using the TS-CAN model for rPPG estimation, provides insight into the model’s decision-making process, demonstrating that the identified regions contribute to the accurate extraction of blood volume pulse signals. In the majority of the analyzed scenarios, the SNR_{raw} and SNR_{max} values were significantly higher when using the Grad-CAM mask (m_2) compared to the reference face mask (m_1). Grad-CAM based masking provides the greatest benefit in stationary conditions and controlled movements. Moderate improvements occur for slow movements and speech, while fast head movements see no advantage or even slight degradation in SNR_{max} . Large head rotations benefit moderately but remain affected by motion artifacts. The analysis of the attention mask window size revealed that SNR_{raw} improves only marginally, while SNR_{max} increases by over 3.2 dB in categories 2, 3, and 4. The proposed approach indicates that integrating an attention module which leverages multiple temporal windows could further enhance the TS-CAN model.

References

1. Verkruyse W, Svaasand LO, Nelson JS (2008) Remote plethysmographic imaging using ambient light. *Opt Exp* 16(26):21434–45
2. Chen X, Cheng J, Song R, Liu Y, Ward RK, Wang Z (2019) Video-based heart rate measurement: recent advances and future prospects. *IEEE Trans Instrum Meas* 68:3600–3615
3. Lewandowska M, Ruminski J, Kocejko T, Nowak J (2011) Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity. In: Federated conference on computer science and information systems (FedCSIS), Szczecin, Poland, pp 405–410. <https://ieeexplore.ieee.org/document/6078233>
4. Wang W, den Brinker AC, Stuijk S, de Haan G (2017) Algorithmic principles of remote PPG. *IEEE Trans Biomed Eng* 64:1479–1491
5. Ni A, Azarang A, Kehtarnavaz N (2021) A review of deep learning-based contactless heart rate measurement methods. *Sensors* 21(11):3719. <https://doi.org/10.3390/s21113719>
6. Cheng C-H, Wong K-L, Chin J-W, Chan T-T, So RHY (2021) Deep learning methods for remote heart rate measurement: a review and future research agenda. *Sensors* 21(18):6296. <https://doi.org/10.3390/s21186296>
7. Selvaraju RR, Cogswell M, Das A et al (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 128:336–359. <https://doi.org/10.1007/s11263-019-01228-7>
8. Yang X, Wu B, Sato I, Igarashi T (2019). Directing DNNs Attention for Facial Attribution Classification using Gradient-weighted Class Activation Mapping. *arXiv*, abs/1905.00593
9. Park S, Wallraven C (2021) Comparing Facial Expression Recognition in Humans and Machines: Using CAM, GradCAM, and Extremal Perturbation. *arXiv*, abs/2110.04481
10. Xiao M, Zhang L, Shi W, Liu J, He W, Jiang Z (2021) A visualization method based on the Grad-CAM for medical image segmentation model. In: International conference on electronic information engineering and computer science (EIECS) 2021, pp 242–247
11. Chien J-C, Lee J-D, Hu C-S, Wu C-T (2022) The usefulness of gradient-weighted CAM in assisting medical diagnoses. *Appl Sci* 12(15):7748. <https://doi.org/10.3390/app12157748>
12. Przybyło J, Kańtoch E, Jabłoński M, Augustyniak P (2016) Distant measurements of plethysmographic signal in various lighting conditions using configurable frame-rate camera. *Metrol Meas Syst* 23:579–592
13. Tulyakov S, Alameda-Pineda X, Ricci E, Yin L, Cohn JF, Sebe N (2016) Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2396–2404
14. Ruminski J, Kwasniewska A, Szankin M, Kocejko T, Mazur-Milecka M (2019) Evaluation of facial pulse signals using deep neural net models. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), Berlin, Germany, pp 3399–3403. <https://doi.org/10.1109/EMBC.2019.8857839>
15. Zhao P et al (2024) Toward motion robustness: a masked attention regularization framework in remote photoplethysmography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops, pp 7829–7838

16. Liu X et al (2022) rPPG-Toolbox: Deep Remote PPG Toolbox. *Neural Information Processing Systems*
17. Liu X, Fromm J, Patel SN, McDuff DJ (2020) Multi-Task Temporal Shift Attention Networks for On-Device Contactless Vitals Measurement. *arXiv*, abs/2006.03790
18. Yu Z, Li X, Zhao G (2019) Recovering remote photoplethysmograph signal from facial videos using spatio-temporal convolutional networks. *CoRR*, abs/1905.02419. <http://arxiv.org/abs/1905.02419>
19. Chen W, McDuff D (2018) DeepPhys: video-based physiological measurement using convolutional attention networks. *CoRR*, abs/1805.07888. <http://arxiv.org/abs/1805.07888>
20. Yu Z, Shen Y, Shi J, Zhao H, Torr PHS, Zhao G (2021) PhysFormer: facial video-based physiological measurement with temporal difference transformer. *CoRR*, abs/2111.12082. <https://arxiv.org/abs/2111.12082>
21. de Haan G, Jeanne V (2013) Robust pulse rate from chrominance-based rPPG. *IEEE Trans Biomed Eng* 60(10):2878–2886. <https://doi.org/10.1109/TBME.2013.2266196>
22. Lin X (2014) Using blood volume pulse vector to extract rPPG signal in infrared spectrum. Master thesis, Eindhoven University of Technology. <https://research.tue.nl/files/47000435/785066-1.pdf>
23. Spetlik R, Franc V, Cech J, Matas J (2018) Visual heart rate estimation with convolutional neural network. In: *Proceedings of british machine vision conference*, pp 1–12
24. Stricker R, Müller S, Gross H-M (2014) Non-contact “video-based pulse rate measurement on a mobile service robot”. In: *Proceedings of 23st IEEE international symposium on robot and human interactive communication (Ro-Man 2014)*, Edinburgh, Scotland, UK. IEEE, pp 1056–1062
25. Sörnmo L, Laguna P (2005) EEG signal processing. In: Sörnmo L, Laguna P (eds) *Bioelectrical signal processing in cardiac and neurological applications*. Academic Press, pp 55–179. <https://doi.org/10.1016/B978-012437552-9/50003-9>

Artificial Intelligence in Healthcare



Custom Attention in Transformer Architectures for Multiclass ECG Beat Classification

Elia Pattacini^{1,2}(✉) and Vaidotas Marozas² 

¹ Polytechnic University of Milan, Milan, Italy
elia.pattacini@mail.polimi.it

² Kaunas University of Technology, Kaunas, Lithuania

Abstract. Electrocardiogram (ECG) beat classification plays an important role in diagnosing and managing cardiovascular conditions. This study introduces a method for classifying ECG beats into normal, supraventricular, and ventricular types using a transformer-based architecture with a custom attention mechanism. The preprocessing steps include filtering, resampling, normalization, and segmentation of ECG signals. Two classification approaches are proposed: (a) a cascade method employing sequential binary classifications, and (b) a multi-class method for direct classification. Synthetic data simulating supraventricular and ventricular arrhythmia is generated to address data imbalance and enhance model performance. The training and testing processes leveraged various datasets, including the MIT-BIH Arrhythmia, Supraventricular Arrhythmia, European ST, and Long-Term ST databases. The results highlight the superior performance of the cascade model, achieving a macro-average F1 score of 0.92 and a precision score of 0.90 on the Long-Term ST database. These findings show the importance of dataset selection and enhancement, demonstrating the model's potential for clinical applications.

Keywords: Electrocardiogram · Arrhythmias · ECG beat classification · synthetic data · RR interval normalization

1 Introduction

Cardiovascular diseases remain a leading cause of morbidity and mortality worldwide, underscoring the need for advancements in early detection and treatment strategies. Electrocardiogram (ECG) analysis plays an important role in monitoring cardiac health, particularly in the classification of heartbeat types to diagnose and prevent severe outcomes. Among these, distinguishing between normal beats (N), supraventricular beats (S), and ventricular beats (V) is critical due to their association with various cardiac pathologies. Figure 1 illustrates typical ECG signals for these three beat types.

The N beats follow the physiological conduction pathway from the sinoatrial node (SA) through the atrioventricular node (AV) to the Purkinje fibers. S beats originate near the atria or the AV node, often lacking P waves and causing compensatory pauses. V beats arise in the ventricles, bypassing normal pathways and presenting widened QRS complexes. Ectopic beats are key indicators of cardiac health, and their frequency is

associated with the risk of future heart disease and events. Studies have shown that a higher frequency of ventricular ectopic beats after myocardial infarction correlates with increased cardiac mortality [1]. Supraventricular beats also carry significant risks, with a high frequency associated with a markedly increased risk of stroke [2].

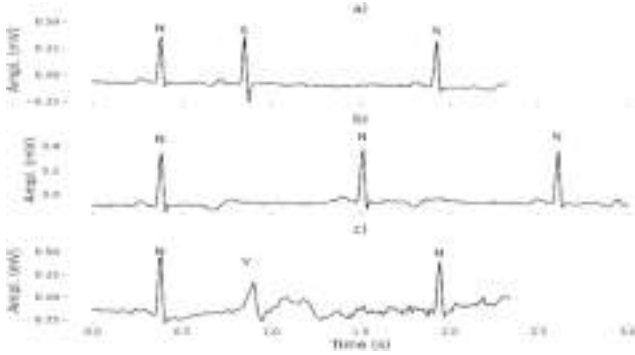


Fig. 1. Supraventricular S (a), normal N (b), and ventricular V (c) beats

The traditional methods for classifying heartbeat types often rely on machine learning models applied to preprocessed ECG signals. However, recent advances in deep learning, particularly transformer-based architecture, have demonstrated superior capabilities in handling temporal data like ECG signals. Transformers, characterized by their attention mechanisms, provide a robust framework for capturing both local and global dependencies in time-series data.

This study introduces a novel transformer architecture with custom attention mechanisms and data augmentation techniques to improve the generalizability and performance of the model. Using advanced data division strategies and using a synthetic dataset, this research aims to address the challenges of class imbalance and variability between data sets. The findings highlight the potential of these methods to improve diagnostic precision, reduce false positives, and contribute to the development of efficient cardiac monitoring systems.

2 Materials and Methods

2.1 Datasets

This study utilizes four publicly available datasets from Physionet, alongside an artificially generated dataset. Table 1 shows an overview of the datasets used. In the following the MIT-BIH Supraventricular database will be referred to as “Short-Term database” while the Long-Term ST Database as “Long-Term database”.

For the training and validation of the model, we utilized two different datasets; for the training, we used the Long-Term datasets while for the validation, the short-term. In addition, we conducted an experiment using synthetic data, training and validating the model exclusively on this generated dataset and evaluating its performance on the

Table 1. Database description.

Dataset	Number of patients	ECG duration (min)	Sample frequency (Hz)
MIT-BIH Supraventricular Arrhythmia	78	30	128
Long-Term ST databases	80	1440	250
European ST	79	120	250
MIT-BIH Arrhythmia (DB2)	26	30	360
Artificial dataset	200	30	1000

real-world dataset. The testing was carried out on all available datasets, ensuring that subsets were taken from different groups of patients than the one used to train the model to prevent bias. The MIT-BIH Arrhythmia Database subset (DB2) was specifically used for testing due to its prevalence in research, covering patient IDs: 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234. Table 2 shows the description of the testing datasets.

Table 2. Testing dataset description.

Dataset	Number of patients	Total number of beats
Short-Term	7	5388
Long-Term	9	37206
European ST	79	49063
DB2	26	44258

2.2 Methods

Block diagram of the method is shown in Fig. 2. Before passing the time series signal as input to the model, a series of preprocessing operations such as lead selection, filtering, resampling, normalization, and beat segmentation was performed. The selected lead displayed a positive QRS morphology, characterized by the expected sequence of Q (downward deflection), R (upward deflection), and S (downward deflection) waveforms. The ECG signals were preprocessed as described in [3]. Baseline wander was mitigated using a dual-stage median filtering process: a 200-ms filter removed QRS complexes and P-waves, followed by a 600-ms filter to suppress T-waves.

The resulting signal, representing baseline wander, was subtracted from the original ECG to obtain a baseline-corrected trace. A 12-tap low-pass FIR filter with a 35 Hz cutoff was then applied to remove powerline noise and other high-frequency interference, ensuring a clean ECG signal. The signals were resampled at 128 Hz and normalized to

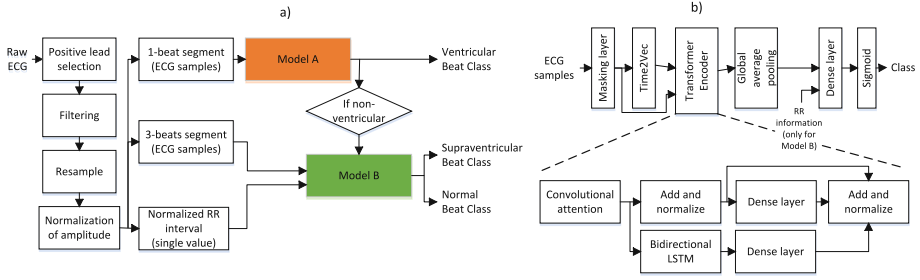


Fig. 2. Block diagram of the method proposed for classification of ECG beats as normal, supraventricular, and ventricular (a), architectures of the Model A and Model B (b)

[0,1] using min-max normalization. Beat segmentation was performed around R peaks, with segment lengths standardized (220 samples for single-beat and 660 samples for three-beat windows). Heart rates outside the 35–250 bpm range were excluded to ensure computational reliability. For efficiency, only 20% of normal beats were used, reduced to 10% in long-duration datasets. Additional information was extracted from the R-R intervals, which were then normalized against a 30-min average. For edge cases, the analysis window was adjusted forward or backward to maintain a consistent reference, accommodating variability in heart rhythms.

To enhance robustness, five augmentation techniques were then applied: 1) amplitude scaling ($0.8\text{--}1.2\times$), 2) quantization (20 levels), 3) drift ($0.1\text{--}0.5\text{ mV}$), 4) noise addition (0.05 mV standard deviation), and 5) time warping. Augmentations targeted ectopic beats more frequently to address data imbalance, with two techniques applied randomly per sample.

Beat classification was approached with a cascade of two binary classifiers: Model A differentiated V beats from others, while Model B distinguished the remaining N and S beats. Both models shared a transformer encoder architecture [4] based on the TRANS-BLSTM-2 [5] network and incorporated a bidirectional long short-term memory (BLSTM) layer within each transformer block.

It also included a custom convolutional attention mechanism inspired by S. Li et al. [6]. This mechanism captures local dependencies in time-series data while enhancing computational efficiency. Figure 2 shows the cascade structure as well as the model's architecture. These models shared the same architecture but differed in the input data while Model A took as input a single beat ECG segment, Model B took as input the three beat ECG segment.

Synthetic ECGs were created using Bachi et al.'s generator [7]. Signals (30 min, 1000 Hz) had simulated conditions like atrial fibrillation, bigeminy/trigeminy, and VPBs, along with noise artifacts such as motion and baseline wander.

2.3 Evaluation Methods

Precision (also called Positive Predictive Value) measures the likelihood that a positive classification made by the model corresponds to a true positive instance:

$$P = \frac{TP}{TP + FP} \tag{1}$$

Recall is defined:

$$R = \frac{TP}{TP + FN} \tag{2}$$

The F1 score combines precision and recall, evaluating the model’s ability to identify minority-class instances while minimizing misclassifications:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

Macro averaging computes performance metrics equally across all classes, making it well-suited for imbalanced datasets.

3 Results

3.1 Synthetic Data

Table 3 presents the performance of the model trained and validated exclusively on synthetic data and tested on the Long-Term dataset. Training with synthetically generated data has yielded promising results, although there is a noticeable decline in performance metrics compared to the model trained on real data (see Table 4 for comparison).

Table 3. Testing results on the Long-Term dataset.

Beat Type	Precision	Recall	F1 score
N	0.812	0.902	0.832
S	0.374	0.532	0.434
V	0.861	0.571	0.702

3.2 Clinical Data

Table 4 presents the classification performance across the four clinical datasets for the testing subsets. The classifier consistently achieved high F1 scores for the N and V categories across all datasets, exceeding 0.87 and 0.83, respectively. Notably, the Long-Term dataset demonstrated the strongest results, with F1 scores of 0.968 and 0.964 for the N and V categories, respectively. For the S category, precision was relatively

high in the Long-Term (0.729) and Short-Term (0.758) datasets but significantly lower in the European (0.321) and MIT-Arrhythmia (0.307) datasets. These results suggest that while the model effectively minimizes false positives for the S category in some datasets, it struggles in others, particularly the European and MIT-Arrhythmia datasets. The weighted averages indicate strong overall performance, with high precision for the N and V categories offsetting the variability in S classification.

Table 4. Testing results of the model on clinical datasets.

Dataset	Category	Precision	Recall	F1 Score	Number of samples
Long-Term	N	0.992	0.945	0.968	18906
	S	0.729	0.971	0.833	5148
	V	0.996	0.934	0.964	13152
	Macro Avg	0.906	0.950	0.955	37206
Short-Term	N	0.899	0.937	0.917	2753
	S	0.758	0.595	0.667	2380
	V	0.800	0.873	0.835	1455
	Macro Avg	0.819	0.802	0.803	5388
European ST	N	0.996	0.873	0.939	44406
	S	0.321	0.751	0.455	1788
	V	0.451	0.933	0.608	3215
	Macro Avg	0.589	0.852	0.663	49063
DB2	N	0.993	0.925	0.958	38708
	S	0.307	0.746	0.435	1094
	V	0.736	0.972	0.837	4456
	Macro Avg	0.679	0.881	0.743	44258

4 Discussion

The model achieved strong precision across various datasets, particularly for the V beat class, even when applied to an independent external dataset. This reflects the model’s capacity for generalization and effectiveness. Although trained and validated on the Long-Term and Short-Term datasets respectively, the model showed good precision with data from patients it had not seen during training, indicating its potential for real-world clinical application. However, a noticeable drop in S beat precision was observed on external datasets not used in training, possibly due to differences in data collection techniques and conditions.

Training a model exclusively on synthetic data has shown promising results, obtaining respectable performances when tested on the Long-Term dataset. For N and V beat types, the model demonstrated good precision and F1 score overall, but it underperformed for S beats. These preliminary findings suggest further research is needed to optimize the use of synthetic data in ECG beat classification.

The cascade method struggles with S beats due to their close resemblance to normal beats, particularly when chest leads exhibit reduced P-wave amplitude. While synthetic data produced strong results for V beats, it may fail to capture the subtle characteristics of S beats without direct distributional comparisons. Databases utilizing torso-mounted modified limb leads (e.g., MIT-BIH Supraventricular Arrhythmia, Long-Term ST) provide clearer P waves, resulting in higher S-beat classification rates compared to those dominated by chest leads (e.g., MIT-BIH Arrhythmia, European ST). An immediate next step is to fine-tune the model using each patient's pre-recorded beats, which should enhance S-beat classification by accounting for individual ECG variations. Additionally, integrating a Grad-CAM algorithm is planned to visualize which segments of each beat receive the most attention, thereby facilitating further improvements based on a clearer understanding of the model's decision-making process.

The proposed model comprises 114,020 parameters (445.39 KB in total) and required 41 min for the first cascade step and 105 min for the second when trained on a dataset of 155,231 samples and validated on 40,048 samples. During inference, it processes approximately 1,300 samples per second, demonstrating efficient computational performance. In contrast, tests with a standard multi-head attention mechanism demanded significantly more computational power, making it impractical for our application and thus not further considered.

Edge computing processes data near its source, e.g., on wearables or smartphones, instead of a central cloud server. This reduces latency, enhances privacy, and improves real-time decision-making. In ECG beat classification, it enables on-device analysis for faster, secure diagnosis without constant internet access. Given the moderate parameter count, the proposed model potentially could be implemented on an edge device.

5 Conclusions

This study demonstrates the effectiveness of transformer-based architecture with a custom attention mechanism for ECG beat classification, achieving strong performance with a cascade model. The use of synthetic data to address imbalance and the superior results on the Long-Term ST database demonstrate the model's potential for clinical applications. Future work can further refine the approach to enhance accuracy and robustness in real-world settings.

Acknowledgment. The authors would like to thank A. Petrénas and L. Mainardi for constructive discussions and feedback.

References

1. Moss A, Davis H, Decamilla J, Bayer L (1979) Ventricular ectopic beats and their relation to sudden and non-sudden cardiac death after myocardial infarction. *Circulation* 60(5):998–1003

2. Huang B et al (2017) Relation of premature atrial complexes with stroke and death: systematic review and meta-analysis. *Clin Cardiol* 40(11):962–969
3. de Chazal P, O’Dwyer M, Reilly RB (2004) Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans Biomed Eng* 51(7):1196–1206
4. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5999–6009). Neural Information Processing Systems Foundation (2017)
5. Huang, Z., Xu, P., Liang, D., Mishra, A. K., Xiang, B.: TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding. arXiv preprint [arXiv:2003.07000](https://arxiv.org/abs/2003.07000) (2020)
6. Li, S.L., et al.: Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Neural Information Processing Systems*, abs/1907.00235 (2019)
7. Bachi L et al (2023) ECG modeling for simulation of arrhythmias in time-varying conditions. *IEEE Trans Biomed Eng* 70(12):3449–3461



Neurosymbolic AI Approach for Dry and Wet AMD Classification Using OCT Images

Aleksandar Miladinović¹(✉), Alessandro Biscontin¹, Miloš Ajčević²,
Agostino Accardo², Dario Marangoni³, Daniele Tognetto³,
and Leandro Infrerra³

¹ Institute for Maternal and Child Health IRCCS “Burlo Garofolo”, Trieste, Italy
aleksandar.miladinovic@burlo.trieste.it

² Department of Engineering and Architecture, University of Trieste, Trieste, Italy

³ Department of Medicine, Surgery and Health Sciences, Ophthalmology Clinic,
University of Trieste, Trieste, Italy

Abstract. Accurate classification between dry and wet age-related macular degeneration (AMD) is crucial for determining appropriate treatment strategies and improving patient outcomes. However, deep learning convolutional neural networks (CNNs) often face challenges when working with small datasets, particularly in the context of rare pathologies, which can hinder their robustness and generalizability. To address these issues, we employed a neurosymbolic approach that integrates medical knowledge in the form of symbolic AI, enhancing the model’s interpretability and reasoning capabilities. The aim of this study was to improve the classification of retinal conditions, specifically dry AMD, wet AMD, and healthy retinas. Our results demonstrated an overall accuracy of 93%, indicating the effectiveness of this methodology in accurately classifying retinal diseases. These findings suggest that the neurosymbolic approach holds promise for advancing diagnostic support in ophthalmology while providing a transparent decision-making framework.

Keywords: Neurosymbolic AI · Optical Coherence Tomography · OCT · Retinal imaging

1 Introduction

Neurosymbolic learning represents a significant advancement in artificial intelligence, particularly in the medical domain, where decision-making processes must often mirror human reasoning. This approach integrates the strengths of deep learning, specifically convolutional neural networks (CNNs) and Deep Learning (DL), with symbolic reasoning, thereby creating a more interpretable and transparent framework for classification tasks.

Until today, DL is the most common type of architecture used for the classification and management of various medical conditions due to its ability to identify pathological features from images effectively [1,2]. Specifically, such algorithms serve as robust tools for the automatic identification and quantification of retinal abnormality signs observable in optical coherence tomography (OCT) images [3–5]. Recently, a variety of ML models have been developed and extensively utilized for analyzing OCT images from patients suffering from significant ocular conditions, age-related macular degeneration (AMD), central serous chorioretinopathy, including diabetic retinopathy, epiretinal membranes, and glaucoma [6–16].

In terms of OCT image classification, the most commonly employed convolutional neural network (CNN) architectures, such as VGG, ResNet, Inception and their derivatives, have demonstrated highly promising performance [17–21]. For example, a study conducted by Infrerra et al. reported impressive accuracy rates (between 93% and 99%) when identifying conditions like epiretinal membranes, intraretinal fluid, and macular neovascularization [22].

While CNNs excel at detecting patterns and abnormalities, they often fall short in providing a transparent model [23]. This limitation can be particularly pronounced when dealing with rare conditions, where CNNs may struggle due to small datasets and insufficient training examples. In contrast, the neurosymbolic framework allows for the integration of medical knowledge in the form of rules, which enhances the model’s performance and explainability of the model [24]. This is particularly important in ophthalmology, where certain retinal conditions may present infrequently, making it challenging for standard CNNs to learn effectively from limited examples [25] and incorporating additional medical knowledge into the model may enhance its performance.

Accurate differentiation between dry AMD and wet AMD is crucial for effective patient management and treatment planning. Wet AMD is characterized by the development of macular neovascularization (MNV), which can lead to rapid vision loss due to fluid accumulation and subsequent retinal damage [26]. In contrast, dry AMD typically progresses more slowly and is associated with the presence of drusen (D) and retinal pigment epithelium changes, which may not lead to immediate vision impairment but can result in geographic atrophy over time [27]. By analyzing the presence or absence of specific retinal abnormalities, clinicians can accurately determine the final pathology. This method aligns with the principles of neurosymbolic learning, where integrating domain-specific knowledge allows for more informed decision-making and improved diagnostic outcomes [22,28].

Therefore, the aim of this study is to demonstrate how the innovative integration of deep learning and symbolic reasoning (neurosymbolic approach) can enhance the classification of dry AMD, wet AMD, and healthy retinas, not only improving model performance but also increasing explainability and interpretability, thereby providing deeper insights into the decision-making process.

2 Methods

This retrospective observational study was conducted at the University Eye Clinic in Trieste and adhered to the ethical standards set forth in the Declaration of Helsinki. The study's protocol was approved by the Regional Ethics Committee (CEUR) of Friuli Venezia Giulia, Italy, under protocol number 17094/2022. All participants gave their informed consent for the use of their data.

In this study, we conducted a retrospective assessment of anonymized optical coherence tomography (OCT) images acquired through an A-line scanning protocol with a measurement span of 9.0 mm. The imaging was performed using the Spectralis OCT apparatus (manufactured by Heidelberg Engineering, located in Heidelberg, Germany), which functions at a central wavelength of 815 nm. This system provides an axial resolution of 3.9 μm per pixel, a lateral resolution of 5.7 μm per pixel, and generates images with dimensions of 768×496 pixels. Our dataset comprised both horizontal and vertical line scans that were focused on the fovea, sourced from participants aged between 18 and 95 years, and collected during the time frame from January 2017 to September 2022.

Finally, 1407 OCT images of dry AMD, wet AMD affected and healthy retinas (469 for each class) were collected.

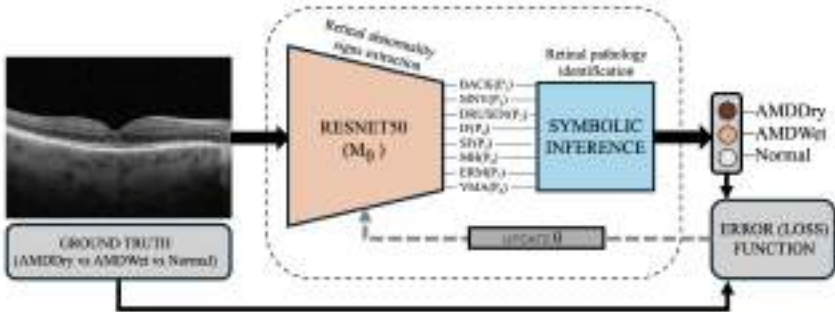


Fig. 1. Illustration of the learning process in the neurosymbolic framework.

2.1 Neurosymbolic AI Reasoning

In this study, for the neurosymbolic AI reasoning, we employed the Scallop programming framework [29]. Scallop allows an integration of logical rules (i.e. expert knowledge) with subsymbolic machine learning approaches (i.e. neural networks, in our case CNN), in a structured manner. This capability is particularly beneficial in medical applications, where domain-specific knowledge can significantly enhance the interpretability and performance of AI decision support systems.

Scallop utilizes provenance tracking to support differentiable reasoning by associating weights, probabilities, or gradients with facts and rules. For our implementation (see Fig. 1), we utilized Differentiable Max-Min Probability Provenance (DMMP), designed to support gradient-based optimization in probabilistic reasoning systems. Unlike traditional provenance systems, DMMP employs dual-number representation, where each provenance tag is augmented with a derivative component. This allows the system to track changes in probability values while enabling efficient backpropagation in neurosymbolic learning models [30].

2.2 Expert Knowledge Representation

In Scallop, expert knowledge is represented in the form of logical rules. These rules, defined by ophthalmologists, encapsulate the symbolic reasoning behind the classification of dry AMD, wet AMD, and Healthy retinas. For this work, we define the following rules that capture domain-specific knowledge:

$$\begin{aligned} \text{DryAMD}((P_3 \wedge \neg P_4 \wedge \neg P_2) \wedge (P_3 \wedge \neg P_5 \wedge \neg P_2) \wedge (P_1 \vee P_7)):- \\ \text{BS}(P_1), \text{ MNV}(P_2), \text{ D}(P_3), \text{ IF}(P_4), \text{ SF}(P_5), \\ \text{ MH}(P_6), \text{ ERM}(P_7), \text{ VMA}(P_8). \end{aligned} \quad (1)$$

$$\begin{aligned} \text{WetAMD}(((P_3 \wedge P_4) \vee (P_3 \wedge P_5) \vee (P_3 \wedge \neg P_7)) \wedge (P_2 \wedge P_1)):- \\ \text{BACK}(P_1), \text{ MNV}(P_2), \text{ D}(P_3), \text{ IF}(P_4), \text{ SF}(P_5), \text{ MH}(P_6), \\ \text{ ERM}(P_7), \text{ VMA}(P_8). \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Healthy retina}(\neg P_1 \wedge \neg P_2 \wedge \neg P_3 \wedge \neg P_4 \wedge \neg P_5 \wedge \neg P_7):- \\ \text{BACK}(P_1), \text{ MNV}(P_2), \text{ D}(P_3), \text{ IF}(P_4), \text{ SF}(P_5), \text{ MH}(P_6), \\ \text{ ERM}(P_7), \text{ VMA}(P_8). \end{aligned} \quad (3)$$

where *AMDDry*, *AMDWet*, and *Healthy retina* represent retinal pathologies; *BACK*, *MNV*, *D*, *IF*, *SF*, *MH*, *ERM*, and *VMA* denote retinal abnormality signs; and P_n represents the probability of the presence of a corresponding retinal abnormality. In particular, the rule (1) for dry AMD is predicated on the identification of Drusen (D) as a hallmark characteristic of this condition. The absence of Intraretinal Fluid (IF), Subretinal Fluid (SF) and Macular Neovascularization (MNV), are more commonly associated with wet AMD. The presence of Choroidal Backscattering (BACK) may occur in later stages AMD [31].

In contrast, the rule (2) for wet AMD is constructed to reflect the condition's defining characteristics, which include the presence of Drusen (D) and Macular Neovascularization (MNV) in conjunction with either Intraretinal Fluid (IF) or Subretinal Fluid (SF) [28].

For the classification of a healthy retinas, the rule is explicitly defined by the absence of all key pathological features [22].

Table 1 presents an illustrative example demonstrating how the rule engine operates using rules (1)–(3) and Differentiable Max-Min Probability Provenance

(DMMP) for classification. The table showcases how different probabilities of retinal abnormality signs influence the classification outcomes for dry AMD, wet AMD, and Healthy retinas. In addition, Fig. 2 shows Grad-CAM activation maps from the fine-tuned ResNet50 model, highlighting key retinal abnormality signs. The first column presents the original OCT images, while the second column shows their corresponding Grad-CAM heatmaps, illustrating the model’s attention to specific pathological features. The four abnormalities displayed are intraretinal fluid, macular neovascularization, backscattering, and drusen. The red regions indicate areas of high importance for classification, while blue regions represent lower attention.

Table 1. Illustrative example demonstrating how the inference engine operates using rules (1)–(2) and Differentiable Max-Min Probability Provenance (DMMP) for classification.

P(Retinal Abnormality Signs)								P(Classification Outcome)		
BS	MNV	D	IF	SF	MH	MER	VMA	AMDDry	AMDWet	Healthy
0.80	0.20	0.90	0.10	0.10	0.30	0.80	0.20	0.75	0.42	0.01
0.60	0.70	0.60	0.70	0.80	0.20	0.10	0.20	0.14	0.71	0.00
0.10	0.00	0.10	0.00	0.00	0.10	0.20	0.10	0.08	0.02	0.65

The objective of deep-learning CNNs in a neurosymbolic framework is to estimate the probability of the presence of these retinal signs abnormalities, even though they are not explicitly trained or supervised on them.

2.3 CNN Deep-Learning Architecture

For the deep learning component of our model, we utilized the ImageNet pre-trained ResNet50 architecture. The final layer of the model was modified to provide independent probabilities for each of the eight retinal abnormality signs.

This adjustment allows the model to perform multi-label classification, enabling it to predict the presence of multiple retinal abnormalities simultaneously. The ResNet50 model was fine-tuned to classify dry AMD, wet AMD, and Healthy retinas, rather than relying solely on direct supervision of the retinal signs (see Fig. 1). The gradient of the pre-trained ResNet50 was updated using the symbolic rules defined earlier, which represent the medical knowledge encoded in the system. The total OCT dataset was partitioned into 80% for training, 10% for validation, and 10% for testing to ensure a balanced evaluation of the model’s performance. During training, the following learning parameters were employed: a learning rate of 0.0001, the Binary Cross-Entropy (BCE) loss function, and a total of 15 epochs.

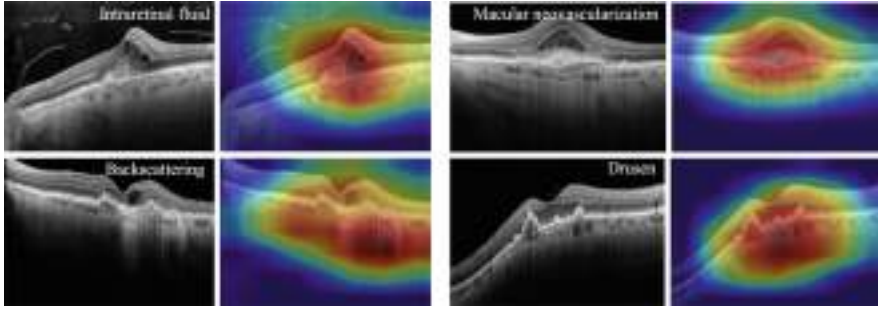


Fig. 2. Example of Grad-CAM activation maps from the fine-tuned ResNet50 model, highlighting key retinal abnormality signs. The first column presents the original OCT images, while the second column shows their corresponding Grad-CAM heatmaps, illustrating the model’s attention to specific pathological features. The four abnormalities displayed are intraretinal fluid, macular neovascularization, backscattering, and drusen. The red regions indicate areas of high importance for classification, while blue regions represent lower attention.

3 Results

The performance of the neurosymbolic learning model was evaluated on the test set. The confusion matrix (Table 2) shows the classification performance for each category: dry AMD, wet AMD, and Healthy controls.

Table 2. Confusion matrix on the test set, expressed as percentages relative to the total number of samples in each class.

Predicted Classification			
	dry AMD	wet AMD	Healthy
dry AMD	1.00	0.00	0.00
wet AMD	0.21	0.79	0.00
Healthy	0.00	0.00	1.00

The detailed classification metrics, including precision, recall, and F1-score, are summarized in Table 3.

The model achieved an overall accuracy of 93% across all classes. The precision for dry AMD was 0.80, reflecting that 80% of the instances predicted as dry AMD were correct. The recall for dry AMD was at 1.00, indicating that all actual dry AMD cases were identified. For wet AMD, the precision was 1.00, while the recall was lower at 0.79, suggesting that some wet AMD cases were misclassified as dry AMD. Healthy cases were classified with precision, recall and F1 of 1.0. None of the pathological retinas were exchanged for healthy ones.

Table 3. Summary of classification metrics, including precision, recall, and F1-score for each class.

Classification Metrics Summary			
Class	Precision	Recall	F1-Score
dry AMD	0.80	1.00	0.89
wet AMD	1.00	0.79	0.88
Healthy	1.00	1.00	1.00

4 Discussion

The implementation of a neurosymbolic approach in this study highlights the critical need for advanced AI methodologies in the medical field, particularly in ophthalmology. Differentiating between wet AMD and dry AMD is essential from a clinical perspective, as these two conditions require distinct treatment strategies and have different prognoses. Accurate classification of these conditions can significantly impact patient outcomes and treatment efficacy.

The neurosymbolic approach combines the strengths of deep learning with symbolic reasoning, allowing for a more robust and interpretable classification system. By integrating domain knowledge into the model, we can create classifiers that are not only accurate but also capable of reasoning in a manner similar to that of an ophthalmologist. This modularized learning process enables the convolutional neural network (CNN) to identify structural abnormalities, even when not explicitly supervised.

The aim of this work was to demonstrate how a neurosymbolic approach can improve the classification of retinal conditions, specifically dry AMD, wet AMD, and Healthy controls, using a dataset of 469 images per class. The results indicated an overall accuracy of 93%, with high precision and recall for each class, underscoring the effectiveness of this methodology.

Moreover, the neurosymbolic approach is particularly advantageous for addressing rare pathologies where datasets may not be large enough to train robust deep learning CNNs. In such cases, the ability to incorporate expert knowledge into the model can enhance its performance and reliability, providing a pathway to better diagnose conditions that are infrequently encountered in clinical practice.

The neurosymbolic innovative approach contrasts sharply with traditional black-box models, which often lack transparency and interpretability. By providing a clear rationale for its decisions, the neurosymbolic model fosters trust and understanding among clinicians, potentially leading to higher adoption rates in clinical settings.

However, a key limitation of this study lies in the simplicity of the defined rules, which may not always capture the full complexity of retinal abnormality signs. While the proposed approach effectively integrates algorithmic supervision

for feature extraction, further validation is needed to ensure that the extracted features are accurately identified.

As this is a preliminary study, future work could explore how refining and expanding the rule set might enhance the localization of retinal abnormalities. More sophisticated rule definitions, such as assigning greater weight to certain signs when detecting, could improve the model's ability to discern pathological features. Additionally, if multiple signs are assigned equal importance for a given pathology, the model might arbitrarily attribute the label to one of them. While this simplification was intentional for the initial study, it remains an area for future refinement.

Overall, this work serves as an initial demonstration of the potential neurosymbolic approaches hold for medical imaging, balancing the advantages of deep learning with enhanced interpretability through symbolic reasoning.

In conclusion, the integration of symbolic reasoning with deep learning through a neurosymbolic framework offers a promising avenue for enhancing diagnostic accuracy in ophthalmology. This preliminary study demonstrates the potential of such approaches to provide more transparent and explainable decision-making processes, ultimately improving patient care and outcomes.

References

1. Jiang F et al (2017) Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2(4):230–243
2. Korot E et al (2021) Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell* 3:288–298
3. Liefers B et al (2021) Quantification of key retinal features in early and late age-related macular degeneration using deep learning. *Am J Ophthalmol* 226:1–12
4. Lee CS et al (2017) Deep-learning based automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Exp* 8(7):3440
5. Schmidt-Erfurth U et al (2022) AI-based monitoring of retinal fluid in disease activity and under therapy. *Prog Retin Eye Res* 86:100972
6. Ting D et al (2017) Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 318(22):2211–2223
7. Gulshan V et al (2016) Development and validation of a deep learning algorithm for the detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316(22):2402–2410
8. Abràmoff MD et al (2016) Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investig Ophthalmol Vis Sci* 57(13):5200–5206
9. Gargeya R, Leng T (2017) Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 124(7):962–969
10. Li Z et al (2018) Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* 125(8):1199–1206
11. Burlina PM et al (2017) Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol* 135(11):1170–1176

12. Grassmann F et al (2018) A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology* 125(9):1410–1420
13. Yousef S et al (2018) Detection of longitudinal visual field progression in glaucoma using machine learning. *Am J Ophthalmol* 193:71–79
14. Lo YC et al (2020) Epiretinal membrane detection at the ophthalmologist level using deep learning of optical coherence tomography. *Sci Rep* 10(1):8424
15. Kim SH et al (2022) Deep learning-based prediction of outcomes following non-complicated epiretinal membrane surgery. *Retina* 42(8):1465–1471
16. Crincoli E et al (2023) New artificial intelligence analysis for prediction of long-term visual improvement after epiretinal membrane surgery. *Retina* 43(2):173–181
17. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. *Computer Science > Computer Vision and Pattern Recognition*
18. He K et al (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp 770–778
19. Szegedy C et al (2016) Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, pp 2818–2826
20. Lee CS, Baughman DM, Lee AY (2017) Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retin* 1(4):322–327
21. Choi KJ et al (2021) Deep learning models for screening of high myopia using optical coherence tomography. *Sci Rep* 11(1):21663
22. Inferrera L, Borsatti L, Miladinović A (2024) OCT-based deep-learning models for the identification of retinal key signs. *J Ophthalmic Res* 1(1):1–10. <https://doi.org/10.1007/s12345-024-00001-0>
23. Solish D (2024) Submacular and intraretinal fluid associated with arteritic anterior ischemic optic neuropathy. *Case Rep Ophthalmol* 15(1):478–482. <https://doi.org/10.1159/000539147>
24. Cennamo G, Montorio D, Carosielli M, Romano M, Cennamo G (2020) Multimodal imaging in choroidal metastasis. *Ophthalmic Res* 64(3):411–416. <https://doi.org/10.1159/000512766>
25. Francis J et al (2021) Clinical and morphologic characteristics of fibroblast growth factor receptor inhibitor-associated retinopathy. *JAMA Ophthalmol* 139(10):1126. <https://doi.org/10.1001/jamaophthalmol.2021.3331>
26. Sheth JU et al (2024) Macular neovascularization. *Surv Ophthalmol*. <https://doi.org/10.1016/j.survophthal.2024.08.003>
27. Lindsley K et al (2016) Interventions for age-related macular degeneration. *Ophthalmology* 123(1):1–12. <https://doi.org/10.1016/j.ophtha.2015.12.004>
28. Miladinović A, Biscontin A, Ajčević M et al (2024) Evaluating deep learning models for classifying OCT images with limited data and noisy labels. *Sci Rep* 14:30321. <https://doi.org/10.1038/s41598-024-81127-1>
29. Li Z, Si X, Zheng L, Stoica I, Gonzalez JE (2022) Scallop: a language for neurosymbolic programming. In: *Advances in neural information processing systems (NeurIPS)*. <https://scallop-lang.org/>
30. Li Z, Huang J, Naik M (2023) Differentiable reasoning with provenance. In: *Proceedings of the ACM on programming languages*, vol 7(OOPSLA), pp 1–27. <https://doi.org/10.1145/3622411>

31. Wheeler J (2024) Treatment for dry age-related macular degeneration: where we stand in 2024. *Curr Opin Ophthalmol* 35(1):1–10. <https://doi.org/10.1097/ico.0000000000001064>



Early Prediction of ICU Readmissions Using Machine Learning in a Colombian University Hospital

Joel Andrés Lerma Pazos¹ , J. E. Camacho Cogollo¹  ,
and Jose Leonardo Mojica Peñaranda² 

¹ EIA University, Envigado, Colombia
javier.camacho@eia.edu.co

² Pablo Tobón Uribe Hospital, Medellín, Colombia

Abstract. Hospital readmissions are common, unplanned, and potentially avoidable occurrences linked to elevated morbidity and mortality rates. In Colombia, the primary concerns regarding ICU readmissions involve the significant expenses associated with treating patients readmitted due to deteriorating circumstances or heightened illness severity, alongside inadequate discharge decisions made by certain physicians or experts, which decrease the efficacy of healthcare facilities. This study aims to develop a predictive model that provides an early warning for the readmission of adult patients to the intensive care unit. In order to accomplish this, a variety of database administration and data analysis technologies were employed, alongside with machine learning models referenced in the literature. These models were trained and validated using data provided by the Hospital Pablo Tobón Uribe (HPTU) in Medellín. As a result of this project, a predictive model was developed with an accuracy score of 0.74 and an AUC value of 0.74, providing valuable information to physicians and specialists when making ICU discharge decisions. This model generates alerts for patients at risk of readmission if discharged, improving decision-making in intensive care settings.

Keywords: ICU Readmission · Machine Learning · Classification algorithms

1 Introduction

Determining a patient's transfer from the intensive care unit (ICU) to the general ward or step-down unit presents a daily problem for intensivists. Extended ICU admissions might elevate the likelihood of infection and healthcare expenses, whereas premature ICU discharge may lead to higher mortality rates and return to the ICU [1, 2]. Therefore, the ICU readmission rate has served as an essential indicator for assessing ICU safety. Despite its importance, a significant number of discharge decisions depend on subjective evaluation owing to the absence of standardized protocols [3, 4].

Developing an objective tool for patient discharge is difficult, due to it must include both the patient's medical status and non-medical criteria. Current research indicates

that ICU readmission rates are affected by factors outside the quality of care, including patient characteristics, length of stay, and various data sources [5].

In the United States, Bice recorded a typical range of ICU readmission rates, varying between 1.2% and 14.5% of discharges [6]. Additionally, certain factors were identified as being associated with a higher risk of ICU readmission, including preexisting disabling neurological deficits and the use of vasoactive drugs during ICU stay. Other patient-related risk factors for ICU readmission include age, baseline condition, comorbidities, disease severity, diagnosis, and organ failure [7].

The readmission of a patient to the ICU can potentially triple the daily cost of ICU care at a healthcare facility, as the patient may present with increased severity of the diagnosed illness due to a lack of knowledge about self-care, therapeutic errors, or premature discharge. The latter accounts for one-third of intensive care readmissions [8].

Developing accurate ICU readmission prediction models based on artificial intelligence tools and identifying physiological variables that predict these clinical events has gained increasing interest in recent years. These advancements assist specialists in making informed decisions for patients and improving the efficiency of high-level healthcare centers [8]. Several studies worldwide have achieved promising results in developing predictive models for intensive care units, either by identifying surgical patients at high risk of ICU readmission [9] or by outperforming existing models in the literature [8]. However, a common characteristic of many of these models is their high degree of customization. That is, their training and parameter tuning were conducted using local data, making them applicable only to the specific population for which they were developed.

Since no studies or research on the development of ICU readmission prediction models have been conducted in Colombia, there is an opportunity to create a model tailored to the characteristics and intensive care units of the country. By leveraging machine learning tools, local clinical records can be analyzed to uncover patterns and identify key variables that facilitate early warning systems. This, in turn, would assist ICU specialists in making better-informed decisions to reduce the number of readmissions.

1.1 Related Work

In recent years, various studies and scientific research have aimed to develop tools, with the help of artificial intelligence (AI), to establish criteria that support medical staff in decision-making and help prevent the health deterioration associated with ICU readmissions. One of the main areas of AI that has recently gained significant popularity in the healthcare field is Machine Learning (ML), which has had a major impact on patient care [10].

Machine Learning gives a computer the ability to make reasonably accurate predictions about the future. In recent years, numerous ML models, such as Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Gradient Boosting, and other ML algorithms, have been used in various studies worldwide to predict potential ICU patient readmissions [11].

One such application was developed by Xue and colleagues, where a regression model was trained to predict the risk of ICU readmission within 30 days after discharge, based on instantaneous measurements, grouped physiological trends, and medication

trends. Although this model slightly improved the area under the ROC curve (AUC) value—0.661 compared to 0.636 in the literature—it exhibited systematic errors in predicting the readmission of patients with longer ICU stays, as it only considered variables recorded in the last 12 h [8].

In a retrospective cohort study, Hammer and colleagues developed a logistic regression model to predict the readmission of surgical patients who required intensive care. The predictor variables were previously identified by the Beth Israel Deaconess Medical Center and Harvard Medical School, MA, USA. The AUC value of the developed logistic regression model was 0.78 (95% CI 0.74–0.82), significantly surpassing the scores reported in the literature. Based on these findings, the study concluded that the evaluated model is an intuitive tool that helps identify surgical ICU patients at high risk of readmission [9]. It is important to note that this model is specific to post-surgical patients and, therefore, cannot be applied to certain cases such as burns, transplants, or pediatric patients.

Another study conducted in an ICU in Spain led to the development of a hierarchical model based on five base Bayesian classifiers using data on demographic characteristics, the Charlson comorbidity index, the admission source and generic syndrome, and the primary reason for admission. This approach identified a top five set of characteristics associated with a higher risk of mortality in the ICU, which include the APACHE II score, the occurrence of cardiopulmonary arrest, patients in an unstable condition due to coma or shock, the admission source, and the primary reason for admission [12].

In 2022, in the Netherlands, De Hond and colleagues conducted a study to evaluate the performance of a model predicting ICU readmission or death within seven days after discharge, before, during, and after retraining and recalibration. The developed model was a gradient boosting algorithm trained on healthcare records from 2004 to 2021. It was later validated using electronic health record data from 2011 to 2019 from a different tertiary care center. The model's performance was assessed using the AUC and calibration metrics (slope and intercept). Subsequently, the original model was retrained and recalibrated, and its performance was evaluated using a temporal validation design. Initially, the model demonstrated moderate discrimination with an AUC of 0.72 (95% CI: 0.67–0.76). However, after retraining, discrimination significantly improved, reaching an AUC of 0.79 (95% CI: 0.75–0.82). With a readmission or mortality rate of 577 out of 10,052 admissions (5.7%) in the new center, the study concluded that external validation and retraining are crucial steps to consider before applying machine learning models to new environments [13].

In conclusion, the use of artificial intelligence, specifically the development of machine learning models, can have a significant impact on medicine and ICU events, providing substantial benefits in decision-making for both patient health and the optimization of hospital resources [10, 14, 15].

In this study, we focused on developing various ML models using different feature engineering techniques to assist clinical specialists in making ICU discharge decisions for their patients based on clinical records from an intensive care unit. The research aimed to identify which predictive model and variables best suit the needs of the Hospital

Pablo Tobón Uribe (HPTU) to generate an early warning system for adult ICU readmissions. The data comes from a private hospital database containing various structured and unstructured variables from adult patients.

The remainder of this article is organized as follows: first, we provide information on the materials and methods used throughout the research, including a general characterization of the data and the machine learning methods applied. Next, we present the obtained results, followed by a discussion of these results in context. Finally, we outline the key conclusions drawn from the findings.

2 Materials and Methods

We conducted a retrospective observational cohort study. The data used was collected between January 1, 2015, and December 31, 2021. The data were provided by the Hospital Pablo Tobón Uribe in Envigado – Colombia. This database includes consent from each registered patient.

2.1 Patients and Electronic Health Records

The data corresponds to cases of adult patients (>18 years) who were readmitted to the ICU within a period greater than two days but less than 30 days, as well as cases of patients who were not readmitted. Each record contains information on administrative, sociodemographic, physiological, and laboratory variables collected during their ICU stay.

2.2 Outcome

The target was named “reingreso”. This variable represents the classification of ICU readmission for each case and categorized into two classes: Class 1 if, the patient is readmitted after ICU discharge within the stipulated period. Class 0, if the patient does not return to the ICU after discharge or if the readmission occurs in less than two days or more than 30 days.

The performance of each model was evaluated using classification metrics: accuracy, to measure the overall proportion of correct predictions; precision, which assesses the proportion of true positives among the cases predicted as positive; recall, which measures the ability to correctly identify positive cases; and F1-score, which combines precision and recall into a harmonic metric. Additionally, the Area Under the Curve (AUC) of the ROC curve was included as a measure of the model’s overall performance in discriminating between classes.

Given the nature of the problem, greater emphasis was placed on the accuracy and AUC metrics, as the target class was balanced. In such cases, these metrics provide a better assessment of each model’s performance and discrimination capability, essential characteristics for considering a model as a good classifier. Furthermore, recall was prioritized over other metrics, as the problem aims to support medical prevention. A higher recall value was favored to correctly predict more positive cases, thereby reducing false negatives. This approach aligns with the clinical objective of avoiding early or erroneous ICU discharges.

2.3 Data Preparation

The model was developed using Python programming language sklearn library. The final number of records was determined through the process shown in the next flowchart. After handling missing data and applying a couple of random class balancing techniques, a final sample size of $n = 1390$ was obtained from an initial sample size of $n = 11,766$. These criteria were followed in other consulted works [5, 16, 17] (Fig. 1).

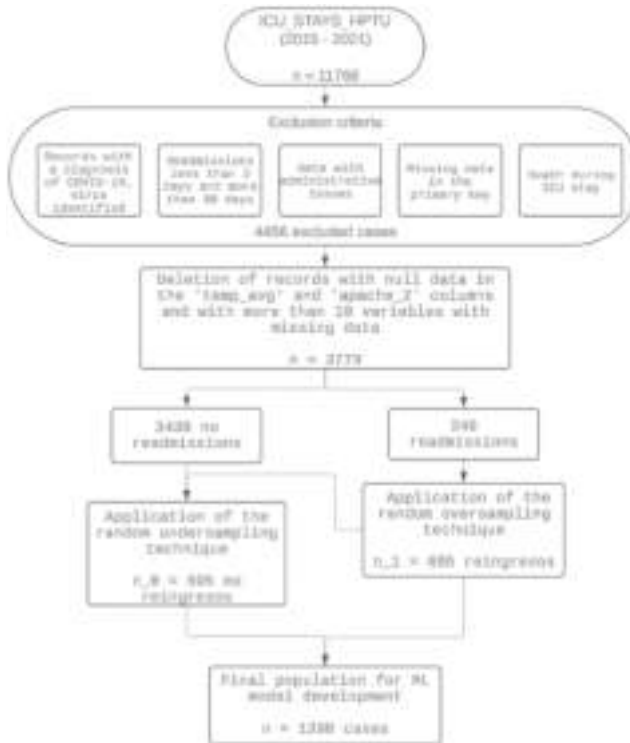


Fig. 1. Flowchart of the final population size.

To improve the quality of the data, adjustments were made to the physiological variables by selecting measurements from the last 24 h of ICU stay. This ensured a standardized time frame for data analysis, given that the length of stay varied across cases and exhibited a high standard deviation (nearly 10 days). Additionally, four statistical metrics were applied during this period: average value, maximum value, minimum value, and standard deviation. A second adjustment was made for laboratory variables. For these features, only the most recent available result was selected, as there was no uniform pattern in the frequency of laboratory tests.

During this preparation process, a column transformer was developed using the 'columntransformer()' tool. The transformation was applied to numerical variables using 'StandardScaler()'. This method adjusts the data: the average = 0 and the variance =

1. For categorical variables, the ‘one-hot encoding’ technique was applied, converting each categorical column into multiple binary columns based on the number of categories present in the variable.

Finally, to handle the remaining missing values, imputation was performed using the average of each variable to preserve the natural characteristics of the data. This approach was chosen because the proportion of missing values per variable was small. The Table 1 shows the proportion of imputed data across the entire dataset.

Table 1. Amount and percentage of imputed data by feature.

Feature	Amount of imputed data	Representative percentage
sum_liq	13	0.93%
glucometria	1	0.07%
sodio	1	0.07%
cant_hco3	1	0.07%
cant_proteina_c	232	16.7%
n_ureico	54	3.88%
lactato	9	0.65%
pcje_fio2	7	0.5%
niv_creatinina	46	3.31%
cant_cloro	146	10.5%
cant_calcio_ion	2	0.14%
cant_neutrofilos	2	0.14%
tco2	5	0.36%
alv_art	5	0.36%

The imputation process was not applied to the XGBoost and LightGBM models, as these models natively support handling missing data.

2.4 Development of Prediction Models and Final Adjustments of the Best-Performing Model

In pursuit of enhancing the model’s classification capabilities, four lists were created, each corresponding to a different group of features. The following table briefly describes the feature groups associated with each list.

The data were randomly divided into three groups: 68% for training, 17% for model validation, and 15% for testing and making final adjustments to the best-performing model. When splitting the data, it was ensured that the proportion of the target class remained consistent across the training and validation sets, maintaining balance.

We developed the following classification models for training and validation of each variable group mentioned in Table 2: Logistic Regression (LR) [18] with Ridge and

Table 2. Description of each group of variables.

Name of group of variables	Description of group of variables
cols_completas	It refers to the entire complete set of features
cols_sin_labs	It refers to the entire complete set of features, excluding laboratory variables
cols_sin_cat	It refers to the set of independent features that only have numerical variables
c_shap	It refers to the set of features that were chosen after applying seven different feature engineering methods

Lasso regularization, Decision Tree (DT), Random Forest (RF) [19], Gradient Boosting Classifier (GBC), XGBoost (XGB) [20], Light Gradient Boosting Machine (LightGBM), and two types of ensembles: Stacking Classifier and Voting Classifier. Some of these models were selected based on a literature review [5, 14, 16, 17].

All these models underwent a hyperparameter search using the GridSearchCV tool, which evaluates all possible combinations of the defined parameters and selects the combination that provides the best performance according to the AUC. This was achieved through five repetitions of cross-validation.

For the last group of variables, seven different feature engineering techniques were applied to select the variables with the highest predictive power. Variables that appeared in at least three of the applied techniques were selected and stored in the list `c_shap`. The techniques used were as follows:

- Significance of the coefficients of the variables with statsmodels.
- ‘feature_importances_’ with Decision Tree model.
- ‘feature_importances_’ with Random Forest model.
- SHAP beeswarm plot.
- Function ‘SelectKBest()’.
- Function ‘forward_selection()’.
- Function ‘backward_selection()’.

A graph was then created to count the number of times each variable was selected by each method. Variables that appeared at least three times were chosen. Using these features, the final variable group was assembled, and each of the previously mentioned ML models was developed.

The model and feature group with the best performance were selected based on the classification metric analysis. With this model, an automated hyperparameter search was conducted using the PyCaret library, which provides functions to automatically optimize a model’s hyperparameters through RandomGridSearchCV() within a predefined search space. The performance of this optimized model was then compared with the version developed using the initial hyperparameters to assess any improvements. Lastly, the test data—unseen by the model—was used to evaluate whether the model had correctly learned how to classify readmissions.

2.5 Managing the Interpretable Algorithm

The prediction process of machine learning models seems like a “black box,” lacking interpretability, without understanding the internal performance and being unable to explain how it reached a certain prediction [21, 22]. This presents a significant issue, particularly in crucial fields like healthcare, where ethical considerations are fundamental. The domain of explainable machine learning is gaining prominence, facilitating the analysis of models and enabling the perception, detection, and comprehension of their decision-making processes, so transforming them into “white boxes” [23]. The understandability of a prediction model is important in clinical practice. The interpretability of a model provides insights into the internal mechanisms of how it works.

The Extreme Gradient Boosting (XGBoost) model is a relatively new tree-based ML algorithm and was developed to improve traditional gradient boosting. Compared to the traditional GBoost model, the 2nd Taylor expansion, the 1st, and 2nd order derivate were used. Besides, regularization to the loss function was added to avoid overfitting due to the complexity of the model.

Shapely Additive explanation (SHAP) is a game theory-based interpretation method of the ML models. It can evaluate both the negative and positive impacts of the input variables on the ML model output. Let a given ML model be trained with the inputs $X_i = (x_1 + x_2 + \dots + x_n)^T$ [20]. SHAP uses an explanation model to evaluate the contribution of each input variable to the ML model.

We combined XGBoost and SHAP to construct an interpretable ML framework, which improved the interpretability of the model and detected important features for ICU readmissions. As a model-agnostic explanation approach, SHAP aids in interpreting predictive models such as XGBoost. It assumes that each feature represents a “contributor” to the predictions of the XGBoost model and assigns them SHAP values; that is, the final prediction can be interpreted as the sum of the SHAP values of all features and the average prediction. SHAP transforms XGBoost’s feature space into a clinical variable space, where each transformed SHAP value corresponds to an original variable. SHAP usually graphically visualizes XGBoost predictions for a better presentation effect.

3 Results

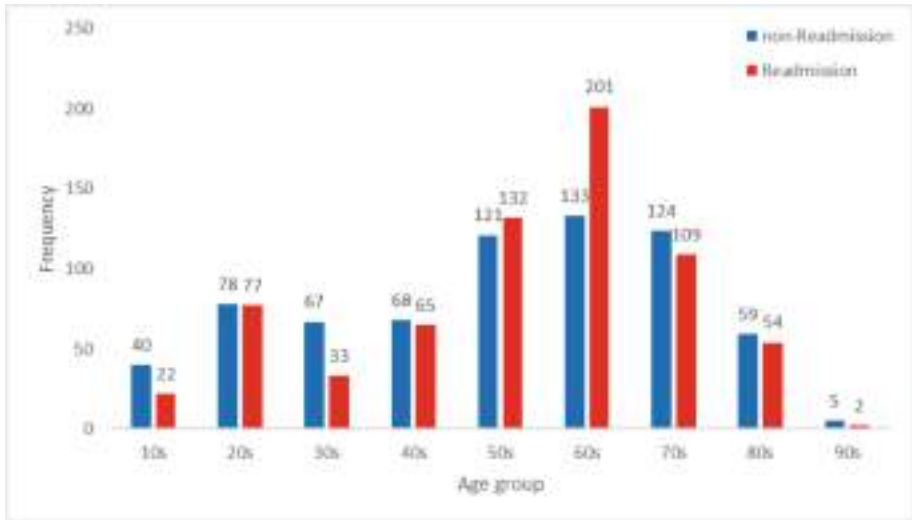
The demographic and administrative characteristics of the studied population are summarized in Table 3, which shows the number of cases analyzed and the number of cases for each target class.

The following figure presents a comparative histogram of age groups differentiated by the target feature. It shows that 50% of readmissions occur among the elderly population, meaning one out of every two readmissions involves a person over 60 years old. Additionally, it is evident that in the 50s and 60s age brackets, the number of readmissions is significantly higher than that of non-readmissions (Fig. 2).

Within the results from each feature engineering method, a matrix was created to cross-reference the features with the results of each variable selection technique, marking the features selected by each specific method. The following figure shows the outcome of this process, where the features appearing in at least three different techniques were selected for the ‘c_shap’ variable group (Fig. 3).

Table 3. Statistics of demographic and administrative characteristics.

Feature	Value
Number of patients	1390
Average age ($\pm$ SD)	55 ($\pm$ 20)
Average days in ICU ($\pm$ SD)	7.7 ($\pm$ 8.2)
Average number of days it takes to readmission ($\pm$ SD)	8.8 ($\pm$ 6.5)
Number of readmissions (Yes / No)	695/695

**Fig. 2.** Histogram of the population's ages differentiated by the target variable.

Additionally, the variable “temp_avg” was included in the final variable group based on the recommendation of the intensivist who guided the study of the medical variables. These were deemed highly important for assessing a patient's condition in the ICU. Table 4 presents and describes the variables selected for the `c_shap` list.

In this work, we tested the ML classification models with each variable group (`'cols_completas'`, `'cols_sin_lab'`, `'cols_sin_cat'`, and `'c_shap'`), and the performance of the three best models from each group was compared. Table 5 presents the results of the classification metrics used to evaluate the performance using AUC, recall, and accuracy. The abbreviations of the models used for building the ensemble models are shown in parentheses.

Table 6 shows the performance achieved with each model developed using the set of features found in the feature engineering process.

Table 7 shows the performance of the selected model after undergoing an automated hyperparameter search using the PyCaret library. The goal of this procedure was intended to enhance the model's performance, nevertheless the results remained worse. Therefore,

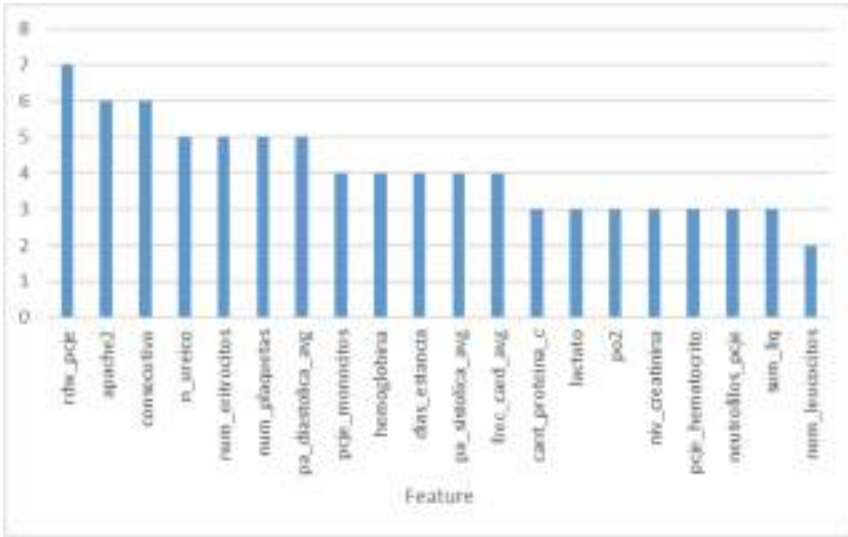


Fig. 3. Frequency of each variable in feature engineering techniques.

the final test continued using the XGBoost model built with the parameters identified through GridSearchCV.

The results of the model test with unseen data are presented in Table 8, which shows the model’s performance on the test data. Graphically, Fig. 4 illustrates the confusion matrix for this model.

One advantage of the tested model is that for this type of classification, patient safety must always be prioritized. As a result, it is critical to reduce classification mistakes in which a readmission is incorrectly anticipated, particularly those in which a non-readmission is predicted only to have the patient readmitted. In this scenario, the model effectively minimizes false negatives, with only 22 cases out of a total of 111 readmissions.

Figure 5 illustrates the impact of each selected variable in predicting the target class, with variables arranged in descending order according to their predictive power. The colors represent the values of each variable, where red indicates high values and blue indicates low values. A possible interpretation of the graph is that the model tends to have a higher probability of predicting a readmission when there are low values for the “num_plaquetas”, “pa_sistolica_avg”, and “po2”, as well as high values for “rdw_pcje” and “frec_card_avg”. This is due to the fact that these factors are most influential when the target variable is classified by the model.

4 Discussion

This study identified the predictive model and clinical variables best suited to the needs of the Pablo Tobón Uribe Hospital to generate early warnings for adult readmissions to ICU. The results demonstrate the performance of various tested Machine Learning

Table 4. Description of the features selected with feature engineering.

Feature	Description	Normal range
‘consecutivo’	Sequential indicator of the admission number	N/A
‘apache2’	APACHE II score, indicator of disease severity	0 – 67 points
‘temp_avg’	Average body temperature	36.1 – 37.2 °C
‘frec_card_avg’	Average heart rate	60 – 100 beats per minute
‘pa_sistolica_avg’	Average systolic blood pressure	90 – 120 mmHg
‘pa_diastolica_avg’	Average diastolic blood pressure	60 – 80 mmHg
‘cant_proteina_c’	Amount of C-reactive protein	Less than 10 mg/L
‘lactato’	Blood lactate level	0.5 – 2.2 mmol/L
‘num_eritrocitos’	Number of erythrocytes in blood	4.7 – 6.1 million/ μ L
‘po2’	Partial pressure of oxygen in blood	75 – 100 mmHg
‘niv_creatinina’	Blood creatinine level	0.6 – 1.3 mg/dL
‘pcje_monocitos’	Percentage of monocytes in blood	2 – 8%
‘pcje_hematocrito’	Hematocrit percentage in blood	38.3 – 48.6%
‘hemoglobina’	Blood hemoglobin level	13.8 – 17.2 g/dL
‘dias_estancia’	Length of stay in ICU in days	N/A
‘neutrofilos_pcje’	Percentage of neutrophils in blood	40 – 60%
‘rdw_pcje’	Red cell distribution width	11.5 – 14.5%
‘sum_liq’	Sum of fluids administered and discarded during the stay	N/A
‘num_plaquetas’	Number of platelets in blood	150000 – 450000/ μ L

models and their potential to serve as a valuable tool for the early identification of adult ICU readmissions. During the experimentation stages, both classical ML models and ensemble models were tested.

Various experiments were conducted with the three datasets to assess the relevance and predictive power of the variables in the classification models. The findings show that variables associated with clinical laboratory results significantly contribute to the detection of possible ICU readmissions. Confusion matrices provided insights into the models’ predictive capabilities, highlighting laboratory variables such as “Red cell distribution width,” “Amount of C-reactive protein,” and “Partial pressure of oxygen in blood” exhibiting high predictive capacity for ICU readmissions. The evaluation results demonstrated the effectiveness of the physiological variables, (such as systolic blood pressure, heart rate, and body temperature) for the performance of the predictive model.

Table 5. Performance of the best-developed models.

Dataset	Model	Accuracy	Precision	Recall	F1-score	AUC
'cols_completas'	GBC	0.8851	0.9197	0.8456	0.8811	0.8854
	XGBoost	0.8919	0.8874	0.8993	0.8933	0.8918
	LightGBM	0.8885	0.9084	0.8658	0.8866	0.8887
'cols_sin_cat'	XGBoost	0.8851	0.9021	0.8658	0.8836	0.8853
	RF	0.8750	0.8783	0.8724	0.8754	0.8750
	LightGBM	0.8682	0.8667	0.8725	0.8696	0.8682
'cols_sin_lab'	RandomForest	0.8277	0.8141	0.8523	0.8327	0.8275
	XGBoost	0.8649	0.8609	0.8724	0.8667	0.8648
	LightGBM	0.8412	0.8269	0.8658	0.8459	0.8411
'c_shap'	XGBoost	0.8903	0.9026	0.8718	0.8869	0.8901
	Stacking classifier (XGB y RF)	0.8860	0.8813	0.8889	0.8851	0.8861
	LightGBM	0.8650	0.8696	0.8547	0.8621	0.8648

Table 6. Performance of each model with 'c_shap' dataset.

Model	Accuracy	Precision	Recall	F1-score	AUC
LR with Ridge	0.6033	0.5950	0.6154	0.6050	0.6035
LR with Lasso	0.6076	0.6000	0.6154	0.6076	0.6077
DecisionTree	0.7553	0.7122	0.8461	0.7734	0.7564
RandomForest	0.8354	0.8047	0.8803	0.8408	0.8360
GBC	0.8397	0.8110	0.8803	0.8443	0.8402
XGBoost	0.8903	0.9026	0.8718	0.8869	0.8901
LightGBM	0.8650	0.8696	0.8547	0.8621	0.8648
Stacking Classifier	0.8860	0.8813	0.8889	0.8851	0.8861
Voting Classifier	0.8860	0.8813	0.8889	0.8851	0.8861

Table 7. Performance of the XGBoost model with hyperparameter search using PyCaret.

Model	Accuracy	Precision	Recall	F1-score	AUC
XGBoost	0.8189	0.8210	0.8174	0.8281	0.8646

Table 8. Performance of the XGBoost model with the test data.

Model	Accuracy	Precision	Recall	F1-score	AUC
XGBoost	0.7836	0.7946	0.8018	0.7982	0.7823

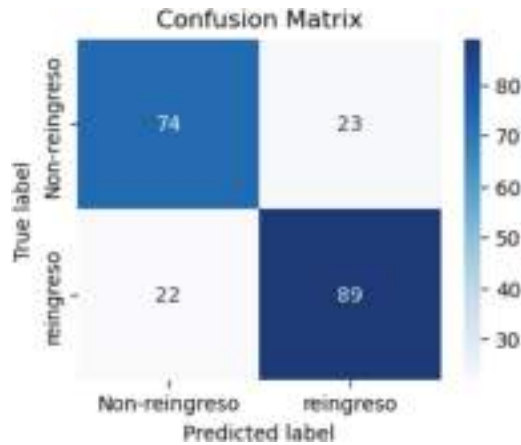


Fig. 4. Confusion matrix of the XGBoost model with the test data.

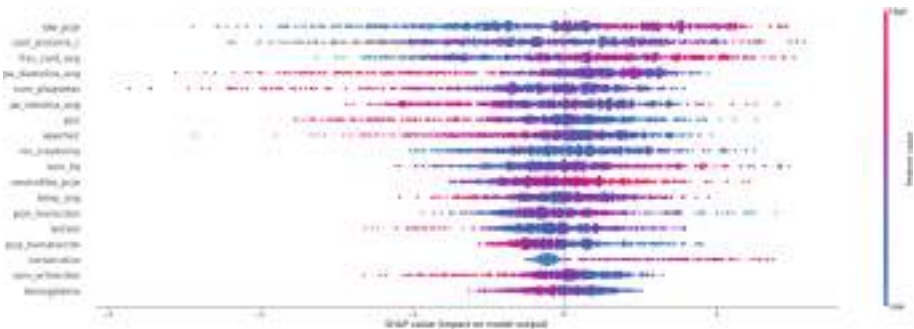


Fig. 5. SHAP values plot of the selected variables.

The analysis of the results reveals the categorical variables such as anisocytosis, poikilocytosis, microcytosis, macrocytosis, stomatocytes, schistocytes, dacrocytes, crenocytes, acanthocytes, and hypochromia—which describe the shape, appearance, or size of red blood cells—have minimal influence on predicting ICU readmissions. The performance of algorithms developed with the variable group excluding these categorical features improved compared to models developed using the ‘cols_completas’ variable group.

The XGBoost models developed with the ‘cols_sin_cat’ and ‘c_shap’ variable groups show similar performance, standing out as the two most robust models among all those constructed. These models achieve the highest accuracy and AUC values, demonstrating superior ability to discriminate between readmission and non-readmission classes. As mentioned earlier, recall is a critical classification metric for this research due to its clinical focus on correctly identifying most true positive cases—specifically, readmissions. Both models yield the highest recall values, indicating their strong ability to correctly classify these cases.

Ultimately, the XGBoost model built with the variable group obtained through feature engineering was selected. This decision was based on achieving similar performance while using 42 fewer features, significantly reducing computational costs despite a slightly lower performance. Notably, the ‘cols_sin_cat’ group includes 60 variables, whereas the ‘c_shap’ group uses only 18 features. The final model tested in our study achieved an accuracy of 0.78 and an AUC of 0.78. This performance is on par with other leading methods. Comparing our results with other machine learning models developed in similar studies, we found our outcomes to be consistent with the current literature (Table 9).

For instance, in the study by Fialho and colleagues, an accuracy of 0.71 and an AUC of 0.72 were achieved for ICU readmission prediction [11]. Similarly, Xue and colleagues developed an alternative model that reached an AUC of 0.66 [5]; while a subsequent study by Hammer reported the validation of a model with an AUC of 0.7 [8]. In a more specific context for Colombia, Serna and colleagues achieved readmission prediction models with an AUC of up to 0.74 [24]. These findings highlight that our model is competitive compared to previous studies and relevant for ICU readmission prediction.

For future studies, it is recommended to have access to a larger dataset, ensuring higher data quality. A limited amount of data for model training can lead to undesirable effects, such as insufficient learning, which may result in the model identifying false patterns and ultimately causing overfitting.

Barriers to Meaningful External Validation

Although the rate of external validation is slowly rising, performance was predominantly worse in external data, questioning the reliability of several recently proposed ML-based outcomes [14]. The external validation appears almost exclusively confined to a few open-source validation sets, most prominently MIMIC [25] and eICU [26]. Nearly 80% of all identified studies conducting external validation utilized a variation of MIMIC. This poses a possible issue, as global studies are predominantly evaluated based on their capacity to maintain performance in patients from the sole US hospital included in MIMIC, which likely does not reflect the broader ICU demographic. This indicates that users and reviewers must meticulously examine assertions of external validation about ICU scoring systems when evaluating products intended for usage beyond the restricted clinical contexts represented by MIMIC. This underscores that although extensive open-source datasets can generate numerous papers in certain domains, they do not inherently enhance the capacity to develop models that generalize, hence constraining their effectiveness in successful clinical translation.

Table 9. comparative table of ML models from the literature.

Name	Year	Database	Population	Model	Features	AUC
Data mining using clinical physiology at discharge to predict ICU readmissions	2012	MIMIC II	25852 cases	Fuzzy model	Physiological and laboratory	0.72
Predicting ICU readmission using grouped physiological and medication trends	2018	MIMIC II	1170 cases	Regression	Physiological and laboratory	0.66
A Tool to Predict Readmission to the Intensive Care Unit in Surgical Critical Care Patients—The RISC Score	2021	MIMIC II	7126 cases	Multivariable logistic regression	Demographics, physiological, surgical factors and laboratory	0.76
Prediction of readmissions, mortality, and infections in the ICU using Machine Learning Techniques	2016	ICU of a hospital in Colombia	53841 cases	Boosting (GBM)	Administrative, demographics and physiological	0.74
Our predictive model to generate an early warning for adult readmission to intensive care units at Pablo Tobón Uribe Hospital	2024	ICU of a hospital in Colombia	1390 cases	Boosting (XGBoost)	Administrative, physiological and laboratory	0.78

The enhancement of external validation through increased variety is obstructed by persistent challenges in obtaining data from various healthcare providers. Apprehensions over data security and privacy deter institutions from disseminating their data. Despite the intention to share, disparities in EHR systems, data quality difficulties, and insufficient semantic compatibility frequently hinder data aggregation without substantial standards and harmonization initiatives. Consequently, it is somewhat unsurprising that external validation disproportionately relies on the limited datasets that are easily accessible. To overcome these obstacles, additional efforts are required to assist researchers in obtaining diverse, multicenter data, encompassing technologies for safe data access.

We support recent initiatives to introduce nuances into the discourse on “clinical validation” [27, 28]. Following this, it is necessary to consider that the data utilized for external validation must be meticulously selected to accurately represent the model’s expected application and particular assertions of generalizability.

5 Limitations

This study also had several limitations. Firstly, the most important limitation of the study is that it is based on retrospective and observational methods. In order to improve the research, prospective investigations are scheduled for future investigations. Secondly, there were other ICU factors that were unable to be included in the analysis due to missing data. Finally, the external validity of the study may have been affected by the fact that the study was performed for one academic Hospital. ML-based classification methods are not without risk, despite their potential. Hospitals frequently differ in the types of patients they see, the services they offer, and the methods they employ to record those encounters. The possibility of “overfitting” is one significant issue in this context. When the system is applied in a different clinical setting, where the initial distinctive features are no longer present, this overfitting may result in predictions that are not accurate. On the other hand, our hospital center comprises 11 distinct ICUs, each employing varied clinical approaches, as they are overseen by different primary team specialties (including infectious disease, neurology, anesthesiology, nephrology, nephrology, pulmonology, surgery, and internal medicine), resulting in a more heterogeneous patient management approach. Our findings suggest that external validation and potentially retraining will likely be required for new implementation sites and machine learning approaches.

6 Conclusions

The final model represents a significant advancement in predicting ICU readmission risk, outperforming traditional scoring systems across diverse Latin American patients cohorts. This study explored the characteristics of the ICU patient database from Pablo Tobón Uribe Hospital using database management and data analysis tools. A wide variety of machine learning models were trained, followed by the selection of the features with the strongest predictive capacity. Finally, the model with the best performance was validated.

The final model, XGBoost, demonstrated good performance, with precision and accuracy metrics indicating an inverse balance between the classes studied. On one hand, the high recall value (0.8018) suggests that the model significantly reduces the number of false negatives. However, its lower precision score (0.7946) indicates that it does not adequately minimize false positives. Regarding the AUC and accuracy metrics, the values of 0.7823 and 0.7836, respectively, suggest that the model has an acceptable capacity to discriminate between classes in each analyzed case.

According to the literature reviewed on the use of models as predictive tools, specifically for ICU readmission, the validated XGBoost model demonstrates reasonable performance in terms of precision and discriminative capacity. This suggests that the model is effective in correctly identifying patients at risk of ICU readmission.

The identified model provides significant added value in the classification of readmissions, as it exhibits a low number of misclassification cases in the false-negative category. This helps minimize the misclassification of readmissions as non-readmissions, thereby preventing erroneous or premature hospital discharge. Ideally, this value should be reduced to zero, as previously explained, since one of the most critical complications leading to patient readmission is premature or incorrect discharge.

The use of computational tools can be highly beneficial for the healthcare sector, as their proper implementation can support healthcare professionals in making critical decisions, such as discharging a patient from the ICU. This, in turn, helps prevent readmissions, improves the quality of healthcare services, and reduces costs.

The extensive range of tools used, the variable analyses conducted, and the results obtained in this research have the potential to serve as a foundation for the implementation of future models and studies focused on using machine learning as a predictive tool in healthcare-related topics. In conclusion, external validation is crucial to evaluate prior to the clinical adoption of a machine learning model in a novel context. Methods such as retraining may enhance model performance at a new location. Clinicians and decision-makers in the ICU should consider this when evaluating the use of new machine learning models in their local environments.

Informed Consent Statement Informed consent was obtained from all subjects involved in the study.

Funding. This research received no external funding.

References

1. Kramer AA, Higgins TL, Zimmerman JE (2016) Can this patient be safely discharged from the ICU? *Intensive Care Med* 42(4):580–582. <https://doi.org/10.1007/s00134-015-4148-8>
2. Woldhek AL, Rijkenberg S, Bosman RJ, van der Voort PHJ (2017) Readmission of ICU patients: a quality indicator? *J Crit Care* 38:328–334. <https://doi.org/10.1016/j.jcrc.2016.12.001>
3. Plotnikoff KM et al (2021) Patient discharge from intensive care: an updated scoping review to identify tools and practices to inform high-quality care. *Crit Care* 25(1):438. <https://doi.org/10.1186/s13054-021-03857-2>

4. Stelfox HT et al (2015) A scoping review of patient discharge from intensive care: opportunities and tools to improve care. *Chest* 147(2):317–327. <https://doi.org/10.1378/chest.13-2965>
5. González-Nóvoa, J.A., et al.: Improving Intensive Care Unit Early Readmission Prediction Using Optimized and Explainable Machine Learning (2023). <https://doi.org/10.3390/ijerph20043455>
6. Bice, T.: ICU Readmissions: Good for Reflection on Performance But Not a Reflection of Quality (2016). United States. <https://doi.org/10.1097/CCM.0000000000001809>
7. Tejerina Álvarez, E.E., Gómez Mediavilla, K.A., Rodríguez Solís, C., Valero González, N., Lorente Balanza, J.Á.: Risk factors for readmission to ICU and analysis of intra-hospital mortality. *Med. Clin. (Barc)* **158**(2), 58–64 (2022). <https://doi.org/10.1016/j.medcli.2020.11.035>
8. Xue Y, Klabjan D, Luo Y (2019) Predicting ICU readmission using grouped physiological and medication trends. *Artif Intell Med* 95:27–37. <https://doi.org/10.1016/j.artmed.2018.08.004>
9. Hammer M et al (2021) A tool to predict readmission to the intensive care unit in surgical critical care patients—the RISC score. *J Intensive Care Med* 36(11):1296–1304. <https://doi.org/10.1177/0885066620949164>
10. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V (2019) Artificial intelligence transforms the future of health care. *Am J Med* 132(7):795–801. <https://doi.org/10.1016/j.amjmed.2019.01.017>
11. Fialho, A.S., Cismondi, F., Vieira, S.M., Reti, S.R., Sousa, J.M.C., Finkelstein, S.N.: Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Syst. Appl.* **39** (2012). <https://doi.org/10.1016/j.eswa.2012.05.086>
12. Delgado, R., Núñez-González, J.D., Yébenes, J.C., Lavado, Á.: Survival in the intensive care unit: a prognosis model based on bayesian classifiers. *Artif. Intell. Med.* 102054 (2021)
13. de Hond AAH et al (2023) Predicting readmission or death after discharge from the ICU: external validation and retraining of a machine learning model. *Crit Care Med* 51(2):291–300. <https://doi.org/10.1097/CCM.0000000000005758>
14. Rockenschaub P et al (2025) External validation of AI-based scoring systems in the ICU: a systematic review and meta-analysis. *BMC Med Inform Decis Mak* 25(1):5. <https://doi.org/10.1186/s12911-024-02830-7>
15. Rajkomar A, Dean J, Kohane I (2019) Machine learning in medicine. *N Engl J Med* 380(14):1347–1358. <https://doi.org/10.1056/NEJMr1814259>
16. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM (2018) Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc* 15(7):846–853. <https://doi.org/10.1513/AnnalsATS.201710-787OC>
17. Markazi-Moghaddam N, Fathi M, Ramezankhani A (2020) Risk prediction models for intensive care unit readmission: a systematic review of methodology and applicability. *Aust Crit Care* 33(4):367–374. <https://doi.org/10.1016/j.aucc.2019.05.005>
18. Mitchell TM, Learning M (1997) Mcgraw-hill science. *Eng Math* 1:27
19. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/a:1010933404324>
20. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
21. Abgrall, G., Holder, A.L., Chelly Dagdia, Z., Zeitouni, K., Monnet, X.: Should AI models be explainable to clinicians?. *Crit. Care* **28**(1), 301 (2024). <https://doi.org/10.1186/s13054-024-05005-y>

22. Fehr, J., et al.: Piloting a survey-based assessment of transparency and trustworthiness with three medical AI tools. *Healthcare* **10**(10) (2022). <https://doi.org/10.3390/healthcare10101923>
23. Band SS et al (2023) Application of explainable artificial intelligence in medical health: a systematic review of interpretability methods. *Inform. Med. Unlocked* 40:101286. <https://doi.org/10.1016/j.imu.2023.101286>
24. Riascos, A., Serna, N.: Predicting Annual Length-Of-Stay and its Impact on Health (2018). <https://pdfs.semanticscholar.org/7582/a9570c45a3636736b10de66d20689eb5ae37.pdf>
25. Johnson AE, Stone DJ, Celi LA, Pollard TJ (2018) The MIMIC code repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc* 25(1):32–39. <https://doi.org/10.1093/jamia/ocx084>
26. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* 5(1):180178. <https://doi.org/10.1038/sdata.2018.178>
27. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA (2020) The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2(9):e489–e492. [https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2)
28. Collins, G.S., et al.: Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384** (2024). <https://doi.org/10.1136/bmj-2023-074819>



Detection of Electronic Protected Health Information in DICOM Objects: Rule-Based Methods vs. LLMs

Dmytro Tkachenko^(✉), Milena Sobotka, Antoni Górecki,
Jakub Kłopotek-Główczewski, Tomasz Neumann, and Jacek Rumiński

FETI, Department of Biomedical Engineering, Gdansk University of Technology,
80-233 Gdansk, Poland

dmytro.tkachenko@pg.edu.pl

<http://www.pg.edu.pl>

Abstract. This study evaluates and compares the effectiveness of rule-based methods and Large Language Models (LLMs) in detecting electronic protected health information (ePHI) within DICOM files, with a particular focus on Polish names and identifiers. The research utilized multiple test datasets containing DICOM and JSON data with identifiers and popular, rare Polish names, as well as false names, which do not exist. We compared several approaches: rule-based methods (STRICT and ANY mode), state-of-the-art (SOTA) models (GPT-4o, DeepSeek v.3, Bielik v.2, and Llama v3.3), and the RoBERTa De-ID Named Entity Recognition (NER) model. Results demonstrated that rule-based methods achieved high precision and recall for structured data, with near-perfect performance for popular and rare names. However, LLMs, while flexible, require high false positive rates and significant processing time. Rule-based algorithms are currently more effective for ePHI detection in DICOM attributes, but they struggle with free-text attribute values. Fine-tuning LLMs is suggested as a potential improvement for ePHI detection in medical imaging data.

Keywords: healthcare informatics · ICT in medicine · DICOM · large language models

1 Introduction

The digital transformation of healthcare has led to an exponential increase in electronic medical data generation and storage. Medical imaging, in particular, represents a significant part of this data, with billions of images generated in healthcare institutions worldwide. While this diversity of data presents new opportunities for medical research and the development of artificial intelligence solutions, it also raises critical concerns regarding patient privacy and data protection. The ability to effectively identify and manage electronic protected health information (ePHI) within medical data becomes more valuable, especially when

such data needs to be shared for research purposes. The ePHI is any information being created, received, stored, or transmitted electronically, which could in any way identify the individual, his health condition, or treatment for the condition [2]. In other words, Protected Health Information (PHI) is personally identifiable information used with one’s physical or mental health (and related health information). In the United States, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule establishes national standards for the protection of individual’s medical records and other personal health information. It establishes rules for the use and disclosure of protected health information (PHI), ensuring that sensitive patient data remains secure while enabling the flow of information required for high-quality healthcare. Compliance with HIPAA involves implementing technical, physical, and administrative safeguards to protect PHI, particularly in electronic forms [9]. Names, addresses, medical record numbers, etc., are presented by HIPAA as identifiers that are considered personal data and should be removed. Precisely, HIPAA defines a limited data set that cannot contain any of the following identifiers [22]: Name; Address (smaller than a state, including street address, city county, and zip code); All elements (except years) of dates, and exact age if over 89); Telephone numbers; Fax number; Email address; Social Security Number; Medical record number; Health plan beneficiary number; Account number; Certificate or license number; Vehicle identifiers and serial numbers, including license plate numbers; Device identifiers and serial numbers; Web URL; Internet Protocol (IP) Address; Finger or voice print; Photographic image - not limited to images of the face; Any other characteristic that could uniquely identify the individual. Similarly, in the European Union, the General Data Protection Regulation (GDPR) provides comprehensive requirements for processing personal health data. The GDPR applies to the processing of personal data across a large range of contexts, including those used in biomedical research and pertaining to vast areas of translational and clinical research activities [11]. The Digital Imaging and Communications in Medicine (DICOM) standard, which is present everywhere in medical imaging, is one of the most important sources of unique challenges for PHI detection and management. DICOM objects store information in structured attributes, ranging from simple fields like Patient Name to complex free-text fields such as Study Description and Series Description. While some attributes have well-defined formats, others allow unrestricted text entry, potentially containing embedded PHI in unexpected locations. The problem of removing annotations stored in image pixel data is quite complex, and it is not addressed in this paper. A particular complexity in DICOM de-identification lies in the requirement to keep certain attributes for technical validity or research purposes. Certain attributes are necessary for a DICOM object to be valid, while others are crucial for linking imaging data to external information, such as clinical metadata, analysis results, or ground truth annotations. This paper aims to compare and evaluate two approaches for detecting PHI in DICOM objects: traditional rule-based methods and modern Large Language Models (LLMs). We propose simple rule-based algorithms based on dictionaries. Next, our research focuses on analyzing

the effectiveness, reliability, and practical implications of both approaches in identifying sensitive information across various DICOM attributes, with particular attention to handling free-text fields where PHI may be less structured or predictable. In Sect. 2, we review related works in medical data de-identification, with a particular focus on DICOM privacy protection and the role of LLMs in privacy-sensitive text analysis. Section 3 describes our methodology, including the creation of our test dataset and the implementation of both rule-based and LLM-based detection approaches. Section 4 presents our experimental results and statistics, while Sect. 5 discusses the implications of our findings and their practical usage. Finally, Sect. 6 concludes the paper with recommendations for future research and potential improvements in ePHI detection.

2 Related Work

Anonymization of medical imaging data, particularly DICOM files, is essential for protecting personally identifiable information (PII) and protected health information (PHI) while ensuring compliance with HIPAA and GDPR. Rule-based methods, which remove personal identifiers from attributes like Patient-Name and PatientID, are widely used. Burbridge [5] highlighted their effectiveness in radiology teaching databases. The Teaching File Transfer Tool (TFTT) integrates anonymization during DICOM extraction from PACS systems and before exporting data to a teaching database. This tool uses the MIRC DICOM Anonymizer to modify, remove, or hash sensitive tags and edit pixel data, ensuring compliance with HIPAA. Shahid et al. [24] proposed a two-stage anonymization framework: initial anonymization during PACS data export and a detailed review of DICOM headers to remove residual identifiers. Muschelli [14] discussed challenges in preserving temporal differences in DICOM headers, crucial for longitudinal analyses. Moore et al. [26] presented a de-identification approach for longitudinal clinical narratives, highlighting challenges in handling multiple time points and indirect identifiers. [13] and Aryanto et al. [3] noted that removing private DICOM attributes can lead to loss of valuable research data, emphasizing the need for tailored deidentification scripts. Rutherford et al. [23] introduced a public DICOM dataset for evaluating anonymization methods. Vcelak et al. [28] addressed burned-in annotations, noting standard tools often fail to detect and remove them. Mishra and Bagade [12] highlighted vulnerabilities in DICOM headers to “man-in-the-middle” attacks, demonstrating encryption methods like TLS and AES enhance security. Recent advancements use large language models (LLMs) like GPT-4 [19], DeepSeek [7], Bielik.ai [4], and Roberta-DeID [20] to analyze unstructured DICOM text. Shiri et al. [25] developed PRIMIS, which anonymizes medical images using sparse coding with ambiguation (SCA) while preserving diagnostic features. Orekondy et al. [18] introduced a system identifying sensitive visual attributes like faces or tattoos, enhancing privacy risk assessment. Abouakil et al. [1] proposed pseudonymization via the PIPE protocol, offering flexible security levels. Sweeney [27] introduced k-anonymity to reduce reidentification risks. Authors in [16] developed a Co-guided Neural Network

(CogNN) for name recognition, improving sensitivity detection with BERTA-based models.

While rule-based methods work well for structured attributes, they struggle with unstructured text and complex data fields. LLMs provide promising solutions. This study evaluates rule-based and LLM approaches on DICOM data with Polish (synthetically generated) names, considering regional factors and linguistic context.

3 Method

In the following section, we describe the datasets used and created for the study, deep learning models adapted in this work, and simple rule-based algorithms. This study assumes the detection of two classes of PHI: last names/surnames and IDs.

3.1 Datasets

A set of datasets was created for the evaluation of handling sensitive data within medical imaging documents:

- **Set IA:** DICOM images with synthetic labels, including last names/surnames and person identifiers.
- **Set IB:** DICOM images with synthetic labels, including last names/surnames and person identifiers, with additional variations such as separators.
- **Set II:** JSON documents containing selected DICOM metadata, where the names of DICOM attributes are provided as a context.

Set IA contains three subsets:

- **Most popular names:** Names selected from the 600 most common names in Poland (Polish SSN registry - PESEL).
- **Rare names:** Names selected from the 600 least common names in Poland (Polish SSN registry - PESEL).
- **False names:** Randomly changed original names as strings not present in the Polish PESEL registry.

Each subset contains 34 DICOM files, resulting in a total of 102 files. Each file includes six simple attributes with names (e.g., PatientName), two free text attributes containing names (i.e., Study/Series Description), and three attributes with IDs (e.g., PatientID).

Set IB contains two subsets:

- **Random names and IDs:** Random values of names and identifiers.
- **Names and IDs with separators:** Names and identifiers with inclusions of single characters from a set of popular separators (e.g., ‘_’;”).

Set II consists of JSON versions of selected DICOM attributes. The JSON files divided also on three categories with popular, rare and random names. Each category includes the following set of keys (selected DICOM attributes' names): SOPClassUID, SOPInstanceUID, StudyDate, SeriesDate, ContentDate, AcquisitionDateTime, ModalityManufacturer, ReferringPhysicianName, PhysiciansOfRecord, PerformingPhysicianName, NameOfPhysiciansReadingStudy, OperatorsName, StudyDescription, SeriesDescription, ManufacturerModelName, PatientName, PatientID, OtherPatientIDs, OtherPatientIDsSequence, PatientBirthDate, PatientSex, PatientAge, PatientWeight, AcquisitionDate. These keys are associated with values that could contain sensitive data, allowing the identification of study participants, time, equipment, or study location. DICOM files use attribute tags (e.g., (0010, 0020)), while JSON files contain the names of the attributes (e.g., PatientID) (Fig. 1). This distinction is important as it provides context for large language models, enabling them to better understand and process the data (attribute's name present the context for the LLMs).

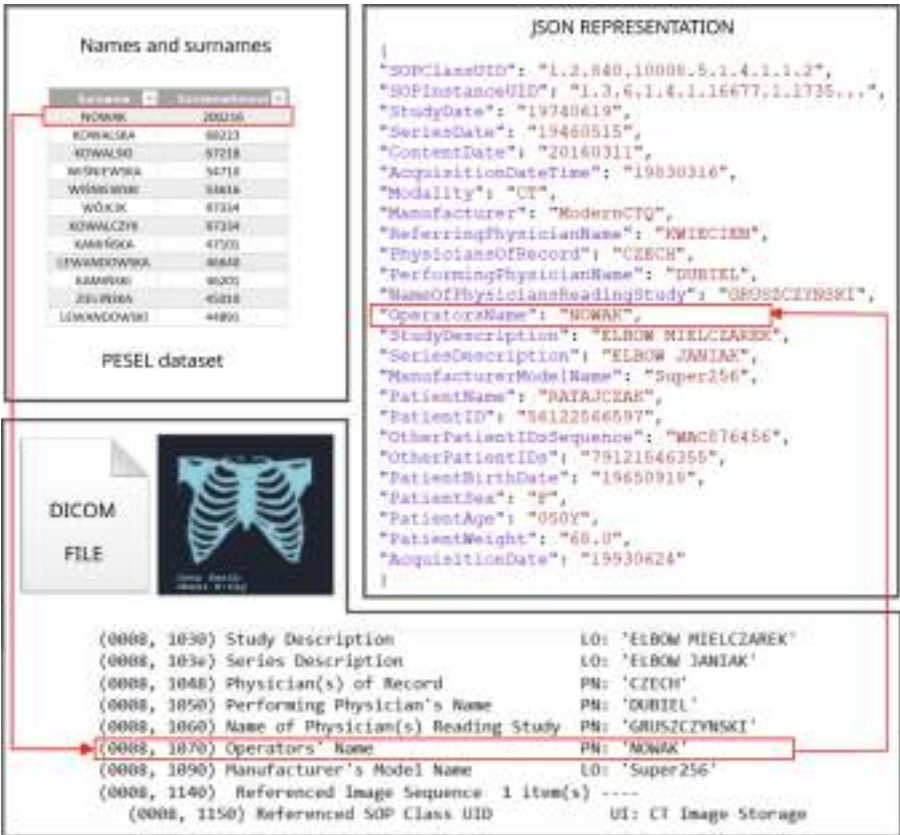


Fig. 1. Data flow using PESEL dataset.

The DICOM files were selected from two public datasets: “A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis” [21] and “COVID19-CT-Dataset: An Open-Access Chest CT Image Repository of 1000+ Patients with Confirmed COVID-19 Diagnosis” [15]. These datasets include DICOM files from various CT scanners manufactured by three different companies. Two first manufacturers appear with 1 model, but the last one has 3 different models, which sums up to 6 CT models overall. The DICOM files were modified using the Python pydicom library, with sensitive data replaced by synthetic values. The modified files were validated using the dicom3tools validator [6], ensuring no structural errors were introduced during the modification process. Each category of DICOM files was accompanied by a pickle file indicating whether the file contained sensitive data (1) or random character strings (0). The datasets were designed to evaluate the ability of large language models to handle sensitive data in medical imaging. The inclusion of various categories of names, identifiers, and metadata ensures a comprehensive evaluation of the model’s capabilities.

3.2 Adapted Large Language Models

Rule-based methods simply rely on predefined patterns, regular expressions, etc., which are used to identify specific types of sensitive data, such as names or identifiers. While well-defined rule-based approaches are highly interpretable and can achieve high precision, they may struggle with variations in data formats. Large Language Models (LLMs) offer a more flexible and context-aware approach to detect sensitive information. Such models can understand the semantics and context of the text, which helps them deal with complex and unstructured data. In our study, we utilized 4 different models to evaluate and compare approaches: GPT-4o [19], DeepSeek v.3 [7], Bielik v.2 [4] (Polish LLM based on Mistral-7B, but its structure is extended) and LLama v3.3 [8] developed by Meta company. Additionally, we adapted and tested a Named Entity Recognition (NER) model, which is called RoBERTa DeID. This model is fine-tuned RoBERTa for the de-identification of medical notes and has been recently used to detect ePHI data for the de-identification procedure presented in [10]. The Bielik 2.0 model was fine-tuned on Polish texts. Many experiments were performed to find the most effective system prompt. The following system prompt in the Polish language was used in experiments described in this study for the Bielik model: *system = ”Jesteś ekspertem w wykrywaniu chronionych informacji w elektronicznych dokumentach medycznych oraz plikach DICOM. Zidentyfikuj wszystkie jednostki, które w ramach wartości zawierają dane typu “data”, “nazwisko”, “identyfikator osoby” występujące w podanym dokumencie JSON. Zwróć wyłącznie wartości jako Named Entities w formacie JSON. Zwróć tylko dokument JSON według wzorca {“nazwiska”: { “wykryty_atrybut”: {“wykryta_wartość”: “poziom ufnosci”}, }, “daty”: { “wykryty_atrybut”: {“wykryta_wartość”: “poziom ufnosci”}, }, “identyfikatory”: { “wykryty_atrybut”: {“wykryta_wartość”: “poziom ufnosci”}, } }. Przy każdym wyniku podaj wartość poziomu ufności. Nie podawaj innych wyjaśnień. ”*

For other models we used the following prompt:

You are an expert in detecting electronic protected information in electronic health documents. Please identify and show all entities (no source code) that contain all values of type ‘date’, ‘surname’ or ‘last name’, and ‘person id’ in the values of the following JSON dictionary: (JSON placeholder) Please return only the values of identified entities in JSON format.

The analyzed text can be divided into two main categories: simple, structured content and free text content. The simple structured content is specific for many DICOM attributes, for example: PatientName, PatientID, Referring-PhysicianName, NameOfPhysiciansReadingStudy, PatientBirthDate, etc. Such simple attributes may be theoretically easy to analyze. However, the problem arises when some de-identification methods generate false, potentially anonymous values for such attributes. For example, substituting a value of ‘Anonymous’ instead of a real patient name for an attribute called ‘PatientName’ can be problematic as the last name ‘Anonymous’ is a legally used name in the USA (USA Census, number of people with this last name is over 100 [17]). The second type of DICOM category represents free text attributes. Some examples include ‘StudyDescription’, ‘SeriesDescription’, and some other attributes. The problem in protected information detection for such attributes can be related to the context of such attribute values. For example, ‘Head’ is a valid last name in the USA (number of people 27292) and in Poland (number of People 17). So, the interpretation of the ‘StudyDescription’ attribute value containing the text ‘Head examination’ can be problematic. A similar situation can be observed for other languages. For example, “Głowa” is a valid last name in Poland, according to the Polish SSN registry, PESEL (the number of males with this name is 784). This name means “head” and the text “Głowa – badanie” means ‘Head examination’. There are types of DICOM attributes that could be more complex including those with data type sequence (SQ) or those with value multiplication (VM) higher than one. Another specific group of complex attributes are private data elements, i.e., attributes that are not covered by the DICOM standard list of attributes but may be introduced into a DICOM object by a company, hospital, or other DICOM object creator. Usually, such private data elements are removed during the de-identification procedure applied to provide a dataset for machine learning purposes. Therefore, we will not analyze such attributes in this study. In this work, we use synthetically generated datasets, containing protected information. We investigate only names and identifiers in different contexts, mainly focused on Polish names and identifiers specific for Polish citizens (PESEL number, ID Card, Passport number, etc.).

In summary, the following models were adapted, focusing on both online and local implementations:

LLMs in online chat mode (original interface of service providers):

- GPT-4o (chat) - <https://chatgpt.com/>
- DeepSeek v.3 (chat) - <https://chat.deepseek.com/>
- Bielik v.2 (chat) - <https://chat.bielik.ai/>

As local implementation: Llama v3.3 - Ollama (<https://ollama.com/>).

Name Entity Recognition models:

- RoBERTa De-ID online version (Hugging Face)
- RoBERTa De-ID implementation in local service (CAISE platform at Gdansk Tech)

3.3 The Rules-Based Method

The simple approach for the rules-based method was designed to analyze last names and a set of Polish identifiers. First, the data dictionary was built with all Polish names from the latest PESEL dataset. All the names were normalized as upper-case strings, converting all characters to ASCII (e.g., changing ‘ą’ into ‘a’). As a result, the ‘dictionary_names’ list is prepared. Next, the algorithm was proposed to analyze the input names. The simplified pseudocode is presented in Algorithm 1

Algorithm 1. Detect Names in DICOM Elements

```

1: procedure DETECT NAMES(path, dictionary_names)
2:   selected_dicom_elements_collection  $\leftarrow$  read_dicom_file(path)
3:   detected_names  $\leftarrow$   $\emptyset$ 
4:   for data_element  $\in$  selected_dicom_elements_collection do
5:     input_name  $\leftarrow$  read_dicom_element(data_element)
6:     names_to_analyze  $\leftarrow$  split_and_normalize(input_name)
7:     for name  $\in$  names_to_analyze do
8:       if name  $\in$  dictionary_names then
9:         detected_names  $\leftarrow$  detected_names  $\cup$  {data_element}
10:      end if
11:    end for
12:  end for
13:  Return detected_names.
14: end procedure

```

The *split_and_normalize(input_name)* subroutine was designed to split the input complex text into tokens. Each string token was converted to upper case format and normalized to ASCII. To split the text, the special characters detected that are used in DICOM and HL7 standards to separate name components, for example, ‘^’, ‘|’, etc. For example, the data type PN (Person Name) from the DICOM standard allows to store name as a set of components, for example “Rumiński^ Jacek Wiesław^^Dr.”. Such text will be converted into a list [“RUMINSKI”, “JACEK”, “WIESLAW”, “DR”]. The presented, simple algorithm version uses a STRICT comparison of names. Another variant of the algorithm was designed to test if any name from the PESEL dictionary is contained in the process name token from the DICOM data element content. The lines 8–12 of the presented pseudocode was appropriately modified:

The first variant of the algorithm was named STRICT, and the second ANY. Special rules were designed to detect official Polish identifiers including PESEL

```

8: for dict_name  $\in$  dictionary_names do
9:   if dict_name  $\in$  name then
10:    detected_names  $\leftarrow$  detected_names  $\cup$  {data_element}
11:   end if
12: end for

```

(Polish SSN), identity card number, passport number, and driving license number. The rules to detect such numbers were implemented according to the official national standards for valid values of such identifiers. The pattern matching technique was used with Python re package. Finally, the rules-based method was evaluated on a dataset containing various of DICOM files with embedded names and identifiers. The evaluation aimed to measure the accuracy, precision, and recall of both the STRICT and ANY variants of the algorithm.

3.4 Prompt Engineering for LLMs for ePHI Detection

It is important to notice that chat-based procedures are sensitive to contexts. Some context-related aspects that can influence the behavior of the LLM chats are: the user, the history of prompts (and responses), etc. Therefore, to investigate typical context of chat usage we invited five different users (co-authors) to create new chat windows (for GPT4o, DeepSeek 3, and Bielik) using their own accounts, computers, etc. The Roberta-DeID model used in two scenarios: using Hugging Face Gradio interface [20] using the downloaded version that was adapted for the CAISE platform (in GitLab-based model repository). The first scenario was used to process JSON documents from test datasets. The second example was used to process DICOM files. The Bielik 2 model was also downloaded from the Hugging Face repository and adapted in Python code to investigate its results when used without a history and for the controlled model parameters, one of which is temperature.

$$ScaledLogits = \frac{OriginalLogits}{T} \quad (1)$$

where T is a temperature.

Temperature is a parameter that controls the randomness of predictions by scaling the logits (unnormalized log-probabilities) before applying the softmax function (Eq. 1). A higher temperature results in more random outputs, while a lower temperature makes the model more deterministic and focused on the highest probability prediction (Eq. 2).

$$P_i = \frac{\exp(ScaledLogit_i)}{\sum_j \exp(ScaledLogit_j)} \quad (2)$$

The models we used were not fine-tuned for Polish names. Therefore, we additionally used Llama 3 model providing the list of Polish names. Llama 3 is a set of foundation models developed by Meta, which natively support multilingual, coding, reasoning, and tool usage [8]. These models are available in various

configurations, differentiated by the number of parameters. For this work, we utilized the 70-billion-parameter version of Llama 3 available on HuggingFace service to evaluate its capability in detecting electronic protected health information (ePHI) within DICOM attributes. We analyzed the model's outputs with and without access to a dataset containing possible Polish surnames. The model was tested in two languages: Polish and English. This will help evaluate the impact of language in ePHI detection.

In the initial phase of the experiment, the model was tested using 10 individual JSON objects from our dataset, each representing DICOM attributes containing false names (i.e., names that do not exist). We test the model only for false names since most of the models have problems with false positives results. After the first step, the model was provided with a list consisting of all possible Polish surnames or last names. The following modification was made to the prompt:

“List of all possible surnames or last names in electronic health documents: <LIST>. Please identify in the following JSON dictionary all entities with names and surnames which are included in the list above:...”

With this added context, the model was again evaluated on the next set of 10 JSON objects.

Finally, we counted all detected ePHI values and compared results within two stages of experiments to evaluate the impact of additional context on LLM ePHI detection accuracy. A similar approach was used in analyzing the NER-based models. JSON files were used for the RoBERTa De-ID online version (Hugging Face), while DICOM datasets were used in the local implementation of the RoBERTa De-ID model. The resources for this study are available at: <https://github.com/dimatkchnk/ephi-detection>.

4 Results

Here, we present the experimental results obtained from evaluating various LLMs and NER models in the context of de-identification tasks.

4.1 LLMs

In Table 1, 2 and 3, results obtained for the application of LLM-based chat systems like ChatGPT4o, DeepSeek 3, and Bielik 2 are presented. Additionally, the results are presented for the Named Entity Recognition model trained using medical (English) documents. The first two tables present True Positive Rate values. The third table presents False Positive Rates results obtained for false names and false identifiers. Additionally, Fig. 2 presents an example of RoBERTa De-ID model output for one of the JSON files. The results are for the reference implementation of the model at Hugging Face platform.

Table 1. Results of the LLM-based detection: TPR values obtained for the test dataset (Set-II) with popular names and identifiers: ↑ - higher better

Model (json)	names (simple) ↑	names (descriptions) ↑	identifiers ↑
GPT4o	1.00	1.00	1.00
DeepSeek 3	1.00	0.72	1.00
Bielik 2	0.997	0.84	0.767
RoBERTa De-ID	0.87	0.86	0.96

Table 2. Results of the LLM-based detection: TPR values obtained for the test dataset (Set-II) with rare names and identifiers: ↑ - higher better

Model (json)	names (simple) ↑	names (descriptions) ↑	identifiers ↑
GPT4o	1.0	0.41	1.0
DeepSeek 3	1.0	0.21	1.0
Bielik 2	1.0	0.21	0.92
RoBERTa De-ID	0.37	0.57	0.98

Table 3. Results of the LLM-based detection: FPR values obtained for the test dataset (Set-II) with false names and false identifiers: ↓ - lower better

Model (json)	names (simple) ↓	names (descriptions) ↓	identifiers ↓
GPT4o	0.789	0.505	0.82
DeepSeek 3	0.815	0.364	0.82
Bielik 2	0.721	0.43	0.66
RoBERTa De-ID	0.436	0.606	0.027

Table 4. Results of R (recall), P (precision) and F1 for the test dataset (Set-II) obtained with the Bielik LLM local implementation: ↓ - lower better, ↑ - higher better

Bielik 2 implementation	names (simple)			names (descr.)			identifiers		
	R	P	F1	R	P	F1	R	P	F1
Popular names and ids ↑	0.983	0.974	0.978	0.230	1.000	0.374	0.707	0.900	0.797
Rare names and ids ↑	0.840	0.996	0.911	0.000	0.000	0.000	0.753	0.800	0.771
False names and false ids ↓	0.832	0.976	0.898	0.000	0.000	0.000	0.593	0.700	0.652

Table 4 presents results obtained for the evaluation of local implementation of the Bielik v2. model. The model was evaluated using a node with NVIDIA-A6000 GPU (48GB RAM) with the following parameters: ‘speakleash/Bielik-11B-v2.2-Instruct’, 4-bits quantization, temperature = 0.1, max_new_tokens = 400.

Table 5 and 6 provide a comparison of the performance of two rules-based algorithms (STRICT and ANY) on the DICOM dataset (Set-IA). Table 7 details the performance of the RoBERTa DeID NER model for the dataset Set-IA.

Table 5. Results of R (recall), P (precision), and F1 for the test DICOM datasets (Set-IA) obtained with the rules-based algorithm 1 (STRICT) ↓ - lower better, ↑ - higher better

Rules-based algorithm 1	names (simple)			names (descr.)			identifiers		
(STRICT)	R	P	F1	R	P	F1	R	P	F1
Popular names and ids ↑	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Rare names and ids ↑	0.997*	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000
False names and false ids ↓	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 6. Results of R (recall), P (precision) and F1 for the test DICOM datasets (Set-IA) obtained with the rules-based algorithm 2 (ANY) ↓ - lower better, ↑ - higher better

Rules-based algorithm 2	names (simple)			names (descr.)			identifiers		
(ANY)	R	P	F1	R	P	F1	R	P	F1
Popular names and ids ↑	1.000	0.773	0.872	1.000	1.000	1.000	1.000	1.000	1.000
Rare names and ids ↑	0.997*	0.765	0.866	1.000	1.000	1.000	1.000	1.000	1.000
False names and false ids ↓	0.887	0.842	0.864	1.000	0.697	0.821	0.000	0.000	0.000

Table 7. Results of R (recall), P (precision), and F1 for the test DICOM datasets (Set-IA) obtained with the RoBERTa DeID NER ↓ - lower better, ↑ - higher better

Rules-based algorithm 2	names (simple)			names (descr.)			identifiers		
(ANY)	R	P	F1	R	P	F1	R	P	F1
Popular names and ids ↑	0.323	0.660	0.434	0.52	1.000	0.684	1.000	0.302	0.464
Rare names and ids ↑	0.263	0.612	0.368	0.052	1.000	0.684	1.000	0.303	0.465
False names and false ids ↓	0.074	0.196	0.107	0.400	1.000	0.571	0.000	1.000	0.000

Table 8. Results of R (recall), P (precision) and F1 for the Set IB DICOM dataset obtained with the RoBERTa DeID NER ↓ - lower better, ↑ - higher better

Rules STRICT	names (simple)			names (descr.)			identifiers		
	R	P	F1	R	P	F1	R	P	F1
Cases with ePHI	1.000	1.000	1.000	1.000	0.368	0.538	1.000	1.000	1.000
Cases with partial names	0.926	1.000	0.962	1.000	0.221	0.362	0.294	0.625	0.400

Table 8 presents the same metrics for NER model as in Table 7, but to the Set-IB.

4.2 LLama 3 Experiment

The experimental results are presented in Table 9, which displays the True Positive Rate (TPR) values. The findings indicate that providing context with names or surnames enhances the detection of false values in descriptive attributes, such

```
{ "SOPClassUID": "1 DATE 2 DATE 840 ID 1000 ID 8.5.1.4.1.1.2", "SOPInstanceUID":  
"1 ID 3.6 ID 1.4 ID 1 ID 16677 ID 1 ID 173557461825501 ID 4000 ID",  
"StudyDate": "19 DATE 390 ID 707 ID", "SeriesDate": "19700 ID 215", "ContentDate":  
"19940 ID 402 ID", "AcquisitionDateTime": "197406 ID 19", "Modality": "CT",  
"Manufacturer": "ModernCTQ", "ReferringPhysicianName": "M STAFF ARCINI AK",  
"PhysiciansOfRecord": "JUR STAFF KI EW STAFF IC Z PATIENT",  
"PerformingPhysicianName": "CZA ID J ROSE K OW PATIENT SKI",  
"NameOfPhysiciansReadingStudy": "MAT PATIENT USZAK", "OperatorsName":  
"R PATIENT ZEPKA", "StudyDescription": "CH PATIENT EST  
ST PATIENT AN LOC ISZ EW PATIENT SKI", "SeriesDescription": "CH PATIENT EST  
KON PATIENT EC PATIENT Z LOC NY", "ManufacturerModelName": "Super ID 256 ID",  
"PatientName": "R PATIENT ADOMSKI", "PatientID": "79121546 ID 355 ID",  
"OtherPatientIDsSequence": "650303314 ID 47", "OtherPatientIDs": "580213311 ID 88",  
"PatientBirthDate": "19340 ID 223 ID", "PatientSex": "F", "PatientAge": "050 Wks Y",  
"PatientWeight": "68.0", "AcquisitionDate": "20000410" }
```

Fig. 2. Example of RoBERTa De-ID model output for one of JSON files. The results presented are for the reference implementation of the model at Hugging Face platform.

Table 9. Results of the Llama 3 with surnames context experiment: FPR values obtained for the test dataset with false names: ↓ - lower better

	names (simple) ↓	names (descriptions) ↓
Prompt in English		
Without context	0.816	0.850
With context	0.533	0
Prompt in Polish		
Without context	0.800	0
With context	0.783	0

as “StudyDescription” and “SeriesDescription.” For these JSON keys, the model returns a TPR value of 0, which is desirable when identifying false names. However, for attributes where valid values can be identified more easily, the model produces a significant number of false positives. Furthermore, the inclusion of contextual information does not substantially improve the results, regardless of the language used in the prompt. Notably, the TPR value for the Polish prompt is lower than that for the English prompt, which may be attributed to the fact that Llama 3 was predominantly trained on English datasets.

5 Discussion

Rule-based methods demonstrated high precision and recall for detecting structured ePHI, especially when the data relies on predefined patterns. The STRICT algorithm achieved perfect recall and precision for popular and rare names, as well as identifiers, indicating its robustness in handling well-defined data. LLMs output a significant number of false positive samples. Moreover, the time needed to process tokens is high, which makes this approach useless in real-world applications, where data stores have a huge number of DICOMs to process.

In many studies, authors focus on TPR results. The presented values are usually high. However, no intensive False Positive study is presented. When FPR is high, it means that the apriori rule can be assumed that all entries contain protected information. This study analyzed how different methods cope with false names (Tables 3 and 9). It was proved that a very high false positive rate characterizes LLMs. The state-of-the-art model RoBERTa De-ID produces many false IDs, treating almost every date as an ID (e.g., see Fig. 3).

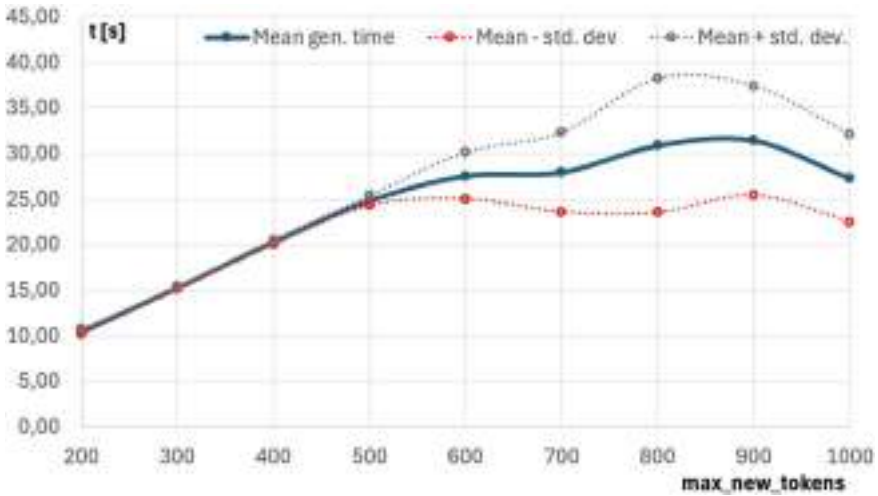


Fig. 3. The total processing time of the Bielik 2 LLM model adapted locally as a function of the maximum number of new tokens (i.e., response tokens, excluding the input prompt).

Recall values for rare names and rule-based algorithms (Tables 5 and 6) are 0.997. Such results show a specific limitation of the proposed rule-based method. The algorithms are based on rules using input tokenization, where a space character is used as a separator. Therefore, when the valid name “B C” is processed, it is divided into tokens B and C, which are independently compared to entries in dictionaries: e.g., “B C”. The dictionary token “B C” is neither equal to “B” nor is contained in “B”. Therefore, it is not recognized and in the Table 5 and

6 we got lower recall metric value. It shows that, for very rare names in Polish registry like “B C”, “B K” or “B.K.” the tokenization is should be modified (which is difficult, e.g., a single letter part of a name).

The study relied on a limited number of test sets, which may not fully capture the diversity of ePHI data in the real world. Most of the LLMs used for the study were not fine-tuned for Polish names and identifiers, which may have affected their performance. Llama local implementation was not able to handle versions of the model with more parameters. For the Bielik Model, the high increase in temperature parameter (more random) influences that some non-name attribute names (e.g., Manufacturer: “ModernCTQ”) were detected as names increasing the FPR. So, the challenge for LLMs for ePHI detection is also an optimal selection of hyperparameters. Fine-tuning LLMs on Polish names and identifiers data, could improve their performance and reduce FPR. Additionally, adding domain-specific knowledge, such as medical terminology, could enhance the model’s ability to detect ePHI in medical documents. Combining rule-based methods with an LLM approach could be useful. The rule-based method will analyze structure data, while LLMs - unstructured. Using bigger models with more parameters, for example, Llama with 70/405 billion parameters, could give different results.

The selected algorithms and methods described in this study were implemented as a Cloud Services as a result of the European IPCEI-CIS project “Cloud Artificial Intelligence Service Engineering (CAISE) platform to create universal and smart services for various application areas”.

6 Conclusion

None of the evaluated LLMs can effectively handle the detection of Polish last names or surnames. The rule-based algorithm is very effective in terms of both recall and precision. However, dictionaries for rule-based algorithms need to be updated when new names appear in registries. Additionally, the input text must be preprocessed to extract proper tokens for further processing. The drawbacks of LLM methods are high FPR values and a long processing time for a single document. The latter is extremely important when thousands of DICOM files should be verified before being used in research studies. This is important since the anonymization methods could not be perfect. We conclude that the rule-based algorithms are more effective for ePHI detection for DICOM-related attributes. First, such methods can accurately detect simple names, even for tokens with special character inclusions. Second, the processing time allows for effectively checking thousands of files in a reasonable time. However, such methods are not flexible regarding the changing nature of new names (e.g., new citizens with new names are registered in the social system), IDs, etc. There are probably no perfect solutions for ePHI detection (i.e., $TPR = 1$, $FPR = 0$). So, the objective of the research criteria should be to minimize the risk but understand that a risk of not detecting ePHI information could exist. The LLMs are very promising and, when fine-tuned with additional data, can limit the FPR. It was shown in this study

for the LLama model. A combination of rule-based and LLMs-based methods can potentially be the best solution for detecting protected health information.

Acknowledgements. The research was supported in part by project “Cloud Artificial Intelligence Service Engineering (CAISE) platform to create universal and smart services for various application areas”, No. KPOD.05.10-IW.10-0005/24, as part of the European IPCEI-CIS program, financed by NRRP (National Recovery and Resilience Plan) funds.

References

1. Abouakil D, Heurix J, Neubauer T (2011) Data models for the pseudonymization of DICOM data. In: 2011 44th Hawaii international conference on system sciences, pp 1–11. <https://doi.org/10.1109/HICSS.2011.136>
2. Alder S (2024) What is ePHI? The HIPAA J. <https://www.hipaajournal.com/ephi/>. Accessed 6 Sept 2024
3. Aryanto K, Oudkerk M, van Ooijen P (2015) Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. *Eur Radiol* 25(12):3685–3695. <https://doi.org/10.1007/s00330-015-3794-0>
4. Ociepa K, Flis Ł, Wróbel K, Gwoździej A, Kinas R (2024) Bielik 7B v0.1: A Polish Language Model – Development, Insights, and Evaluation. [arXiv:2410.18565](https://arxiv.org/abs/2410.18565). <https://arxiv.org/abs/2410.18565>
5. Burbridge B (2020) DICOM image anonymization and transfer to create a diagnostic radiology teaching file. *Int J Radiol Imaging Technol* 6:065. <https://doi.org/10.23937/2572-3235.1510065>
6. Clunie DA (2024) Dicom3tools, Snapshot 20241003103319. DICOM Validator - dciodvfy. <https://dclunie.com/dicom3tools/workinprogress>
7. DeepSeek-AI, Liu A, Feng B, Xue B, Wang B, Wu B et al (2024) DeepSeek-V3 Technical Report. [arXiv:2412.19437](https://arxiv.org/abs/2412.19437). <https://arxiv.org/abs/2412.19437>
8. Dubey A et al (2024) The Llama 3 Herd of Models. [arXiv. abs/2407.21783](https://arxiv.org/abs/2407.21783). <https://doi.org/10.48550/arXiv.2407.21783>
9. Dwyer SJ, Weaver AC, Hughes KK (2020) Health Insurance Portability and Accountability Act: Definitions
10. Haghir H, Baidya R, Dvoretzskii S, Maier-Hein KH, Nolden M (2024) A hybrid AI and rule-based approach to DICOM de-identification. In: MIDI-B challenge workshop, MICCAI
11. Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A et al (2021) General data protection regulation. *Hypertension* 77:1029–1035. <https://doi.org/10.1161/HYPERTENSIONAHA.120.16340>
12. Mishra A, Bagade P (2023) MalDicom: A Memory Forensic Framework for Detecting Malicious Payload in DICOM Files. <https://doi.org/10.48550/arXiv.2303.00045>
13. Moore S et al (2015) De-identification of medical images with retention of scientific research value. *Radiographics* 35:727–735. <https://doi.org/10.1148/rg.2015140244>
14. Muschelli J (2019) Recommendations for processing head CT data. *Front Neuroinform* 13:61. <https://doi.org/10.3389/fninf.2019.00061>
15. Mostafavi SM (2021) COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis. *Harvard Dataverse*, V1. <https://doi.org/10.7910/DVN/6ACUZJ>

16. NameRec*: Highly Accurate and Fine-grained Person Name Recognition [arXiv:2103.11360v2](https://arxiv.org/abs/2103.11360v2) (2021)
17. NameCensus.com. Anonymous last name popularity, history, and meaning. <http://namecensus.com/last-names/anonymous-surname-popularity/>. Accessed 16 Jan 2025
18. Orekondy T, Schiele B, Fritz M (2017) Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. <https://doi.org/10.48550/arXiv.1703.10660>
19. OpenAI (2025) ChatGPT (January 2025 Version): Large language model. <https://chat.openai.com/>
20. Liu Y et al (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. [arXiv:1907.11692](https://arxiv.org/abs/1907.11692). <https://arxiv.org/abs/1907.11692>
21. Li P, Wang S, Li T, Lu J, HuangFu Y, Wang D (2020) A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) [Data set]. The Cancer Imaging Archive. <https://doi.org/10.7937/TCIA.2020.NNC2-0461>
22. Limited data set under HIPAA. <https://www.hipaajournal.com/limited-data-set-under-hipaa/>
23. Rutherford M, Mun SK, Levine B et al (2021) A DICOM dataset for evaluation of medical image de-identification. *Sci Data* 8:183. <https://doi.org/10.1038/s41597-021-00967-y>
24. Shahid A, Bazargani MH, Banahan P, Mac Namee B, Kechadi T, Treacy C, Regan G, MacMahon P (2022) A two-stage de-identification process for privacy-preserving medical image analysis. *Healthcare* 10(5):755. <https://doi.org/10.3390/healthcare10050755>
25. Shiri I et al (2024) PRIMIS: privacy-preserving medical image sharing via deep sparsifying transform learning with obfuscation. *J Biomed Inform* 150:104583. <https://doi.org/10.1016/j.jbi.2024.104583>
26. Stubbs A, Uzuner Ö (2015) Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *J Biomed Inform* 58(Suppl):S20–S29. <https://doi.org/10.1016/j.jbi.2015.07.020>
27. Sweeney L (2002) k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Based Syst* 10:557–570
28. Vcelak P, Kryl M, Kratochvil M, Kleckova J (2019) Identification and classification of DICOM files with burned-in text content. *Int J Med Inform* 126:128–137. <https://doi.org/10.1016/j.ijmedinf.2019.02.011>



The Influence of Color on Semantic Ulcer Segmentation Using Deep Learning Models

Muhammad Usman^(✉) and Jacek Rumiński

FETI, Department of Biomedical Engineering, Gdansk University of Technology,
80-233 Gdansk, Poland
muhammad.usman1@pg.edu.pl
<http://www.pg.edu.pl>

Abstract. Accurate and early image analysis is crucial for proper diagnosis and treatment. Convolutional neural networks (CNNs) have the ability to precisely classify and segment the affected wound areas, which has revolutionized biomedical imaging for automatic detection and diagnosis. However, CNN models often face challenges like generalization and overfitting issues due to limited wound image data. In addition, RGB images are computationally intensive. They may prompt the model to focus on irrelevant features, such as color variations, instead of the most important ones, like silhouettes, ultimately leading the model to the overfitting problem. This study investigates the influence of color on ulcer semantic segmentation. We explore different color-to-grayscale conversion techniques to study the learning behavior of the CNN model. Traditionally, image conversion from an RGB color space to grayscale uses fixed transformation parameters, e.g., YUV color model weights, which are standardized for human observers. The YUV color model separates luminance (Y) from chrominance (U and V), allowing efficient brightness and color information representation. However, machine-based processing does not require ‘seeing’ the content but focuses on extracting the essential features for model training, data augmentation, or future design of task-specific sensors. We designed a universal model architecture to analyze the learning procedure’s ability to learn task-specific weights for color-to-grayscale conversion. The experiment results show how grayscale images based on learned weights could be used in task-specific semantic segmentation compared to color images and grayscale images obtained using traditional conversion techniques. The grayscale images based on learned weights are computationally efficient, and simplified feature representation helps the model emphasize the most relevant attributes while suppressing irrelevant ones. The source code is available on [1].

Keywords: Convolution neural network · Foot ulcer segmentation · Image Modalities · learned grayscale · RGB and gray-scale images

1 Introduction

The amount of money spent on wound care is increasing in the healthcare sector. Approximately 8 million people were injured in 2018, and Medicare costs (USA)

were estimated to be between \$28.1 billion and \$96.8 billion [2]. Wounds that do not heal properly or where the healing process does not restore anatomical and functional integrity after three months are considered chronic wounds. Time is of the essence; the longer a wound is allowed to deteriorate, the more difficult it is to heal, and early diagnosis is the most effective means of reducing wound care costs [3,4].

Chronic wounds are more likely to form in persons who are obese or have diabetes. Diabetes affects 34.2 million Americans and 463 million people globally, and by 2030, this number is expected to increase by 25% [5]. Diabetic foot ulcers (DFUs) are a relatively common form of chronic ulcers on the lower extremities. Regardless of the presence of peripheral vascular disease, neuropathy can diminish or eliminate the sensation of pain in the foot, thus often resulting in diabetic foot ulcers (DFUs), ranging in depth from shallow to deep. An increased mortality rate may result from these wounds or ulcers if they are not adequately managed [6,7].

For appropriate wound care and management, early and precise assessment of the severity of foot wounds can be crucial. Accurate tissue analysis and wound area assessment are essential for proper treatment and diagnosis. Wound attributes such as area, volume, and stage can be identified with the help of precise wound image segmentation [8]. In turn, these characteristics can be used to assess and treat chronic wounds, track the healing trajectory of the wound, plan future interventions, estimate the risk of hospitalization for the patient, or estimate the healing time [9], which can significantly lower hospitalization and amputation rates. It takes a great deal of effort and knowledge for professionals to manually mark the wound area and determine its severity without any inaccuracies. These factors are detrimental to the management of DFU because appropriate wound healing requires prompt and precise wound care decisions and treatments. Therefore, an automated approach is crucial for quick wound assessment, investigation, and treatment. Automatic segmentation of wound images can benefit significantly from computer-aided diagnosis and ailment location [10].

In binary wound segmentation, each pixel in the image is categorized as wound or non-wound. In the beginning, researchers proposed traditional wound segmentation methods involving learning algorithms. For example, methods [11–14] utilize watershed segmentation, threshold-based techniques, feature extraction methods, and similar approaches for semantic wound segmentation. Challenges faced by these methods include sensitivity to lighting conditions, skin color variations, resolution, effective feature engineering, and manual parameter optimization.

Over the years, with the unprecedented success of convolution neural networks (CNNs) in the medical imaging domain, researchers have also introduced deep learning-based segmentation frameworks that outperformed the traditional methods for wound segmentation. A CNN model for diabetic wound segmentation based on MobileNet-v2 was presented in [15]. The prediction masks from MobileNet-2 were converted into binary masks. Subsequently, to create the final segmentation mask, post-processing techniques for hole-filling and noise removal

were applied to the predicted mask. Liu et al. [16] introduced a dual-view segmentation technique that uses a learnable lighting correction module as a preprocessing step. This method employed characteristics from original and illumination-corrected images, enhancing the final result of wound segmentation. Cao et al. [17] proposed a hierarchical segmentation and multilevel classification paradigm to classify diabetic foot ulcers (DFUs) into five grades based on Wagner’s wound assessment criteria.

The above methods face challenges in manual feature engineering or encounter overfitting and generalization problems due to various factors—for example, lack of available training data, skin color variations, and different lighting conditions. In semantic segmentation, especially in the ulcer wound image task, structural and textural information like wound silhouettes are very important. Moreover, the training data for wound images is also limited, and the model focuses on learning the color variations while neglecting the task-specific salient features. As a result, the model struggles with generalization and overfitting problems. Gray-scale images are a simplified representation that has the potential to help the model (e.g., using augmentation) focus on task-specific relevant information while suppressing irrelevant ones.

In [18], authors review 20 studies on measuring wound sizes using image processing techniques. Many of the studies used color-to-grayscale conversion as the first preprocessing step. All these approaches used traditional conversion methods based on fixed weights of transformations between RGB input data and luminance/intensity related color models such as HSV (hue, saturation value) or YUV/YCbCr (Y-luminance, UV or CbCr two chrominance components). These studies, and others presented in the next section, do not analyze to what extent color improves semantic segmentation results compared to single-component images, especially when the grayscale image is obtained not for fixed color-to-grayscale conversion weights but for learned domain-specific weights. It is also challenging to represent colors in digital images relatively constant over varying illuminations and different imaging sensors. Single-component images can potentially reduce the influence of color changes on the subsequent processing steps, and therefore, many image processing tasks use color-to-grayscale conversion. However, reducing the color components is a potential limitation of the information contained in an image. Hence, a need exists to analyze the influence of color and color reduction on wound semantic segmentation.

This study aims to investigate the effect of color on ulcer wounds in semantic segmentation. First, it is investigated to what extent domain-learned color conversion weights can improve the results of semantic wound segmentation compared to traditionally used color-conversion methods based on fixed, YUV color model weights. Second, we aim to analyze how color improves wound semantic segmentation results compared to grayscale images, primarily obtained using domain-learned color conversion weights. To reach these aims, first, we adopt different traditional RGB-grayscale conversion methods to convert the input RGB image to a single component representation. Then, we propose a learned grayscale technique that adaptively maps RGB color images to the learned

grayscale feature maps. We train the semantic segmentation models using various color compositions and compare the results.

The key contributions of our work are as follows:

- a) We designed a simple, universal model architecture to learn the task-specific weights for color-to-grayscale conversion.
- b) We demonstrated that using learned weight for color-to-grayscale conversion leads to much better results than traditional RGB-grayscale conversion methods.
- c) We show that for the best model architecture, no statistical difference was found comparing segmentation metrics obtained for models trained on RGB images and grayscale images converted with learned weights.
- d) We demonstrated that color preprocessing using the adjusted retinex method [19] can improve semantic segmentation results for the applied (not tuned) model for the similar but new domain.
- e) We compared the influence of several color preprocessing steps on the semantic segmentation of ulcer images.

2 Related Work

This section reviews the previous works on wound image segmentation, including deep-learning and traditional machine-learning-based segmentation approaches.

The first step in diagnosing and treating of chronic wounds is to measure the wound area. Several traditional image-processing methods have been applied to classify wound tissue. For example, color descriptors and texture detectors have been used to automate the monitoring of the wound healing process, to extract information from wound images, and to classify skin patches as normal or abnormal [20,21]. However, these experiments did not provide reliable tools for process automation. Likewise, Song et al. [22] delineated 49 attributes employing a feature engineering approach based on traditional machine learning methods, including K-mean clustering, thresholding, region growth in grayscale and RGB, as well as edge detection. Subsequently, the resulting features were fed into a multilayer perceptron (MLP) and radial basis function (RBF) to learn and evaluate the segmentation results. Some other techniques include creating a red-yellow-black-white probability map [23], which is then optimized for thresholding or region growth [24]. However, feature engineering techniques require human involvement in feature selection and cannot adapt to image irregularities, leading to degraded predictive performance. These methods are subject to several constraints, including sensitivity to skin color, the involvement of a certain degree of feature engineering, the lack of full automation of the end-to-end process, and the manual parameter tuning involved. The authors in [25] review different methods to measure wound area, including elliptical estimation, square counting, photogrammetry using imaging devices, and 3D methods such as laser 3D scanners as well as stereophotography and 2-camera systems. Similarly, a study in [26] evaluates the accuracy of digital planimetry (DP) with adaptive

calibration for measuring wound area on curved surfaces. The DP with adaptive calibration exhibits significantly lower error (0.60% vs. 2.65% and 2.23%) and higher precision, eliminating systematic errors, compared to the SilhouetteMobile device and a standard DP. The method is 4.4 times more accurate and 7.4 times more precise, making it one of the best methods for measuring wound area on curved surfaces.

Recent advancements in deep learning have enabled automated feature extraction and data-driven learning. Researchers now leverage convolutional neural networks (CNNs) to identify wound-affected tissues and segment chronic wound areas effectively [15]. To mitigate the above limitation in chronic wound segmentation, Goyal et al. [27] exploit fully convolution networks (FCN) using different model configurations, including FCN-32s, FCN-8s, FCN-16s, and AlexNet. These models were first pretrained in ImageNet [28] and Pascal VOC [29] dataset and then evaluated on the DFU dataset, consisting of 600 wound images. FCN-16 architecture performed adequately, achieving a 79.4% dice score. However, segmentation images with smaller wound areas and irregular borders remained challenging. U-Net being a famous architecture in semantic segmentation tasks, Niri et al. [30] employed it on the ESCALE dataset for diabetic foot ulcer segmentation. In another approach, authors [31] introduced a pre-processing stage before the training of the CNN network to filter noise effects. Later, transfer learning techniques were introduced to elevate the generalization ability of the deep learning model and trained on the Mask-RCNN model to segment chronic wound images [32].

An ensemble network including LinkNet and U-Net employing pretrained EfficientNetB1 and EfficientNetB2 encoders, respectively, with additional pre-training using the Medetec dataset was proposed by Mahbod et al. [33] for the FUSeg challenge 2021 [33]. The segmentation performance was improved through fusion techniques, test time augmentation, and fivefold cross-validation. To constrain boundary information for the automated segmentation of foot wounds, Edge-OCRNet was first presented in [34] and used the ConvNeXt backbone architecture and edge loss function. The Edge-OCRNet segmented mask was then post-processed to enhance the predicted mask. Global and local attributes are equally crucial for accurate segmentation tasks. A WSNet with a global-local architecture was then introduced by Subba et al. [35], which uses the entire image and its patches to extract high-level semantic information and local context in wound images. The model is first trained to classify wounds, and then the segmentation data is used to refine the model further. WSNet achieves a dice score of 84.7% when tested on its own donated dataset, WOUNDSEG, which includes eight different types of wounds.

Color-to-grayscale conversion was used in many approaches focused on wound image analysis. All proposed methods use traditional color-to-grayscale conversion based on HSV, YCbCr, and other color models designed with a human-in-the-loop. In [36], the authors use color-to-grayscale conversion for the reference ruler tick detection and wound area measurement before applying ISO-DATA classification to find a threshold for the image. The combination of RGB

and grayscale images was proposed in [37]. The authors converted the RGB Region-Of-Interest using ITU-R BT.709 luma coefficients. Next, they modeled the marginal distributions of the three main classes of the ROI gray image using a linear combination of discrete Gaussians (LCDG). In [38], the authors used color-to-grayscale conversion before Otsu’s thresholding to obtain the wound segment. The color-to-grayscale conversion was also used [39] before image segmentation. The Authors used grayscale versions of images to apply normalized cuts for segmentation-based detection of wheals on the skin. In [40], the authors selected the saturation plane of the HSV color model to create a single grayscale image used for wound segmentation with the active contour algorithm.

Analyzing research on the processing of wound color images reveals that many approaches use color information in semantic segmentation. However, the existing approaches have not demonstrated the role of color in semantic segmentation. Moreover, all studies that use color-to-grayscale conversion in an initial wound color image processing phase use fixed conversion weights. These weights are established based on human vision properties using color models such as YUV or HSI. So, there is a need to design and perform a study to answer research questions, including the following: 1) Does the machine-learned conversion from color to grayscale enhance semantic segmentation outcomes compared to using fixed conversion weights? 2) To what extent does color information in wound images improve the semantic segmentation results compared to the use of single component images obtained either by the traditional color-to-grayscale conversion using fixed weights or by the use of learned conversion weights?

3 Method

This section describes the proposed approach to investigate the influence of color in semantic segmentation tasks. In particular, we investigate how learned weights for color-to-grayscale conversion can impact the model’s learning ability to focus on specific features with less information than RGB input.

We employed two different U-Net architectures in various configurations to investigate/ evaluate the research question. One model uses a U-Net architecture and a VGG-16 as an encoder, pretrained on the ImageNet dataset. The second model comprises the LiteSeg model using pretrained MobileNet as a backbone. Using pretrained encoders for the initial extraction of features is vital for available datasets with a limited number of cases. Other parts or encoder/decoder architecture are fully trained. Figure 1 presents a block diagram of our architecture. Overall, this architecture consists of four main blocks. First, the input RGB image is passed through a specially designed layer that could be attached to any model to learn the color-to-grayscale conversion weights. This layer is skipped for original or preprocessed RGB inputs. Additionally, we use different color conversion techniques to convert an input RGB image into: 1) Gray World white balance [41], 2) Adjusted retinex white balance [19], 3) grayscale using Luma (L), and 4) grayscale using weights. We followed traditional ITU-R BT weights to convert the RGB image into a grayscale image. After color conversion,

images are fed to the base model. The model's head performs pixel classification for semantic segmentation.

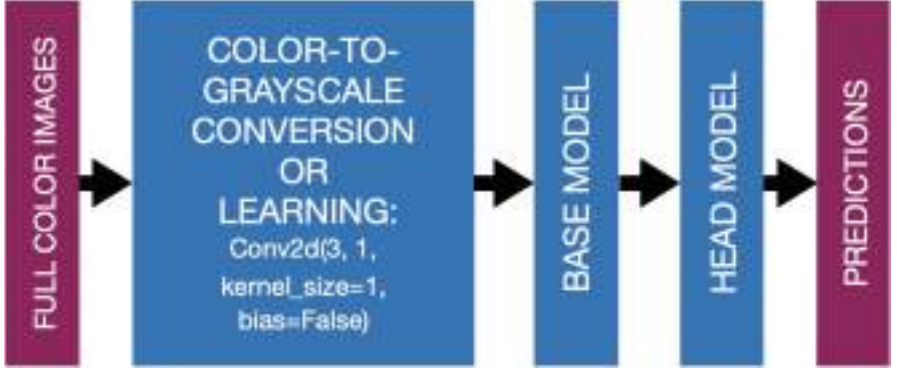


Fig. 1. Description of the proposed method.

3.1 Color Conversion

Luma (Y') or luminance (Y) is often used in many practical computer vision applications as the lightness dimension for color images. Typically, it is calculated as the weighted average of the gamma-corrected color components, i.e., R , G , and B . The weights, as presented in Eq. (1), were formulated to correspond with the cone sensitivity functions in the retina, guaranteeing that grayscale conversion is coherent with human vision.

$$Y' = W_1 \cdot R + W_2 \cdot G + W_3 \cdot B \quad (1)$$

where W_1 , W_2 , and W_3 are the weights corresponding to each RGB color.

As shown in Eqs. (2), (3), and (4), different sets of W_1 - W_3 weights values have been proposed to represent the RGB contribution to perceived lightness, e.g., [42–44]:

$$Y'(Rec.601, SDTV) = 0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B \quad (2)$$

$$Y'(Rec.709, HDTV) = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B \quad (3)$$

$$Y'(Rec.2020, UHD TV) = 0.2627 \cdot R + 0.6780 \cdot G + 0.0593 \cdot B \quad (4)$$

The ITU-R BT.601 weights, as in Eq. (2), are often used as a popular standard for color-to-grayscale conversion (e.g., in the popular Python framework PIL). These traditional color-to-grayscale conversion methods, as in Eq. (2)–(4), were defined for human observers considering the sensitivity functions of cones in the retina. However, a machine has no such limitation, and the input grayscale

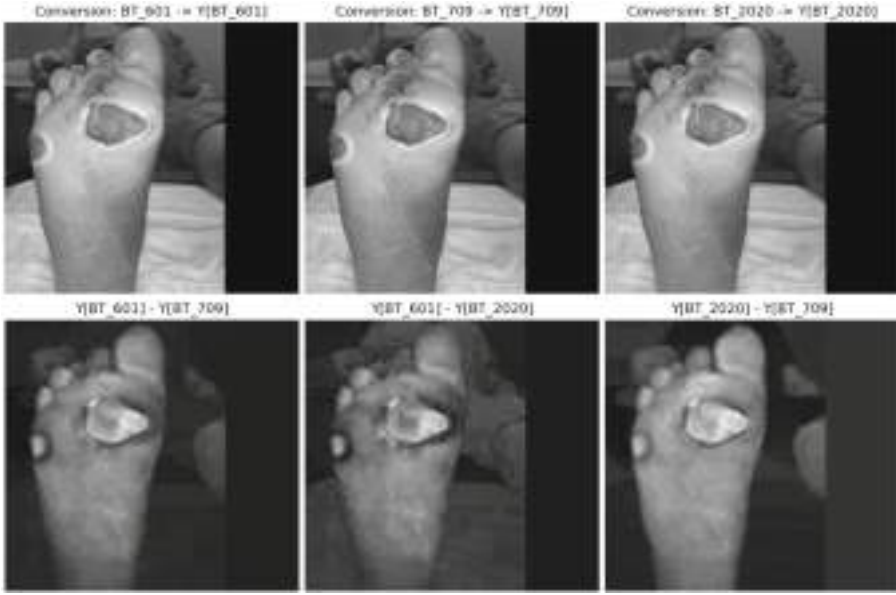


Fig. 2. Top (from left): RGB to grayscale conversion outputs using different ITU-R standards (the PIL.Image.convert('L') method uses ITU-R BT.601). Bottom (from left): differences between some conversion results.

image can be a result of data acquisition using a dedicated, single-channel camera or can be a result of a dedicated preprocessing step aimed at improving the generalization capabilities of a deep learning (DL) model. Figure 4 depicts different grayscale images using different conversion methods (Fig. 2).

In this work, we leverage the ability of CNN to learn data-driven features for learned grayscale transformation from an RGB image to a single color space. For this purpose, a CNN layer is introduced in the color conversion stage, presented in Fig. 1. In deep learning frameworks, the cross-correlation operation is used for convolution (since weights are learned). This operation is defined as in Eq. (5).

$$Y(C_{\text{out}}) = \text{bias}(C_{\text{out}}) + \sum_{k=0}^{C_{\text{in}}-1} W(C_{\text{out}}, k) \cdot \text{input}(k) \quad (5)$$

where C is the number of input/output channels, and W is the learned weights.

Our first layer is defined as Conv2D (3, 1, kernel = 1, bias = False) without activation function. Therefore, this CNN layer receives 3 channels (RGB) at the input with a kernel size of 1 and outputs a single-channel feature map without bias, as presented in Eq. (6)–(7).

$$Y(C_{\text{out}}) = 0 + \sum_{k=0}^2 W(C_{\text{out}}, k) \cdot \text{input}(k) \quad (6)$$

$$Y(C_{\text{out}}) = W_0 \cdot \text{input}(0) + W_1 \cdot \text{input}(1) + W_2 \cdot \text{input}(2) \quad (7)$$

The results is equivalent to (1) with learnable weights W_0, W_1, W_2 . The resultant single-channel feature map could be equivalent to a grayscale image (when a grayscale color map is used), which is optimized for machine learning rather than human visual interpretation. This CNN-based color space transformation enables the model to emphasize vital information and learn low-level features, like texture, edges, contours, and patterns. As a result of this approach, the model can avoid focusing on irrelevant information such as hue variations and extract robust feature representations, resulting in better generalization and less tendency to overfitting. This is especially beneficial when the training data is small.

4 Experiments Settings

In our experimental analysis, we used two publicly available datasets, including Foot Ulcer Segmentation Challenge (FUSeg) [45] and Advancing the Zenith of Healthcare Wound and Vascular Center (AZH-WVC) [15] datasets. Two semantic segmentation models, LiteSeg_MobileNet and UNet_VGG, with pretrained weights, are used as base models in this experiment. First, we train each model using RGB image as input and evaluate their performance on the test dataset; then, we train the model using the color conversion module in the preprocessing stage and compare the accuracy results from both configurations. For color conversion, we train models using each color-to-grayscale conversion method, including learned weights, RGB adj. retinex, RGB adj. retinex learn gray, RGB gray world, RGB gray world learn gray, Luma, and Weights.

4.1 Datasets

Details of the datasets used for training and testing are given below.

The Chronic Wound Dataset [15]: AZH-WVC includes 1109 photos of 889 patients with ulcer wounds. The dimensions of each image are 512×512 pixels. There are 832 photos in the training dataset and 278 in the test dataset.

FUSeg Dataset [45]: The Foot Ulcer Segmentation (FUSeg) dataset comprises 1210 clinical images. 810 images and their corresponding labels are available for training, while 200 images are reserved for validation. The remaining 200 images and their masks were not made public by the challenge organizer. All photos in this collection possess dimensions of 512×512 pixels.

In this study, we used only the FUSeg dataset for training. The evaluation was performed on the FUSeg validation and AZH-WVC test subsets. It is essential

to underline that no models were trained using any image from AZH-WVC to show the generalization properties of the trained models using different color-preprocessing approaches.

Figure 3 presents examples of images from both datasets. Please note that examples from the AZH-WVC dataset are presented as original images (without preprocessing). These images contain a dominant area of zero-padding values, which is challenging for segmentation using a model trained on another dataset.

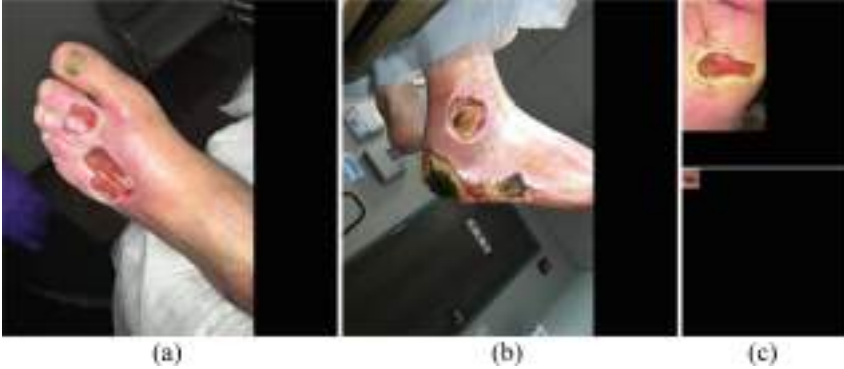


Fig. 3. Examples of wound images from FUSeg and datasets. Images (a) and (b) are from FUSeg, while images in (c) are from the AZH dataset.

4.2 Model Training

Each model with each configuration was independently (random start) trained 10 times. As a result, 80 models (8×10) were obtained for the UNet_VGG architecture, and 80 models (8×10) were obtained for the LiteSeg architecture. For statistical analysis, we used 9 best models for each configuration (one worst was skipped to keep the variance comparable between experiments). Usually, the best models are selected, but to analyze the uncertainty, we used 9 best models to analyze mean values and standard deviations of IoU, Dice, FPR, and FNR metrics.

All experiments were performed using identical training codes (except for differences in model definition) prepared in PyTorch. No augmentation was used. Other training parameters are the Adam optimizer (initial learning rate = 0.001), number of epoches = 60, cross-entropy loss function, and the StepLR learning rate scheduler with gamma = 0.1 and step size = 10.

5 Results

This section illustrates the experiment results and evaluates the segmentation performance. Different performance metrics are considered in this evaluation,

including IoU, Dice, False positive rate (FPR), and False negative rate (FNR). As an additional metric, we calculate evaluation accuracy. It was greater than 94%. However, the accuracy is calculated for both classes (with dominant background), so it is not very informative. Therefore, we report statistics on the above-mentioned segmentation metrics.

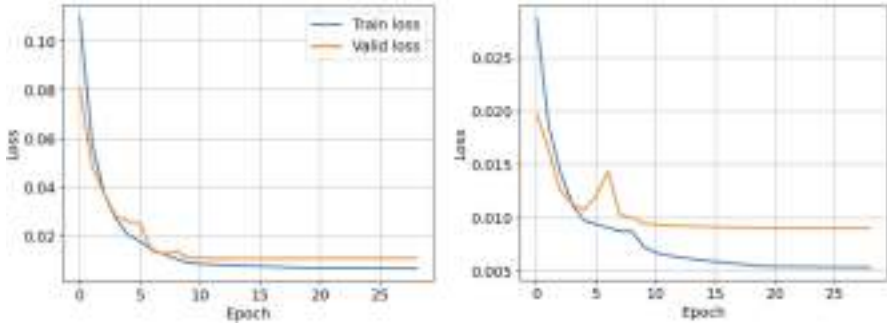


Fig. 4. Example training curves (in blue) and validation curves (in orange) for the UNet-VGG model (left) and LiteSeg model (right) on RGB images preprocessed using the adjusted Retinex method.

Figure 5 presents an example of the qualitative results of wound segmentation from the UNet-VGG model when using learned weights color-to-grayscale conversion. Visual inspection reveals that the model effectively segments the wound image precisely and accurately. While comparing the predicted mask with the actual wound image, we observe that our model with the learned weights module can accurately encompass the contour of the wounded area, which is identical to the RGB image.



Fig. 5. From left: an example of the original photography of a foot ulcer, ground truth mask, and predicted mask using UNet-VGG model with learned weights for color-to-grayscale conversion.



Fig. 6. From left: the result of color image preprocessing (conversion to grayscale using the BT-601 standard), the output of the first layer of the UNet_VGG model (Conv2D with kernel size = 1 and no bias), the output of the color image conversion using weights of the first layer with the overlaid, predicted ulcer segment.

Figure 6 plots three different images after the color conversion process. The image on the left is after color-to-grayscale conversion using the BT-601 standard, while the image in the middle is extracted from the output of the learned weights module of the UNet-VGG model. Similarly, the image on the right is produced using the learned weights (extracted from the first layer) and manually applied to the RGB image. This shows how three learned weights, i.e., W_1 , W_2 , and W_3 (Eq. 1), can be applied after the knowledge obtained from the machine learning process. The last image demonstrates the segmented ulcer overlaid on the new grayscale image. It is noticeable that these new grayscale images explicitly highlight the shape of the affected area and delineate a distinct boundary line of the wound.

We further analyze the significance of each color conversion method for the given RGB input by extending our evaluation procedure with quantitative data from our experiment findings. For this evaluation, we include UNet_VGG and LiteSeg_MobileNet deep learning architectures. Both models are trained on the FUSeg dataset and tested on the FUSeg validation subset and on the AZH test dataset.

Table 1 and 2 show the testing results on the FUSeg and AZH datasets, respectively. It is noticeable that the model achieves significantly lower results using traditional, fixed conversion weights to the RGB image. However, the learned grayscale image still achieves comparable IoU and Dice scores, while the training parameters for the learned grayscale image are three times less than those for the RGB image.

We performed tests to analyze the statistical difference between results for DICE metrics obtained for selected groups. For example, the results obtained for 1) RGB input images and 2) grayscale images converted with the learned weights were not statistically different for the FUSeg validation set (p -value = 0.052 for the Mann-Whitney test, Table 1) and statistically different for the AZH test set (but p -value = 0.042 for Mann-Whitney test).

Table 1. Performance comparison of UNet_VGG on FUSeg dataset (G - grayscale)

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.769	0.071	0.868	0.048	0.041	0.014	0.091	0.059
G: learned weights for RGB	0.732	0.030	0.845	0.020	0.052	0.008	0.141	0.098
RGB: adj. retinex	0.785	0.015	0.880	0.009	0.051	0.008	0.088	0.047
G: l. weights adj. retinex	0.695	0.056	0.819	0.041	0.043	0.012	0.163	0.065
RGB: gray world	0.777	0.009	0.875	0.005	0.045	0.008	0.087	0.012
G: l. weights gray world	0.723	0.022	0.839	0.015	0.055	0.019	0.171	0.175
G: Luma	0.373	0.101	0.536	0.100	0.067	0.096	0.452	0.084
G: Weights	0.445	0.055	0.614	0.053	0.068	0.037	0.346	0.077

Table 2. Performance comparison of UNet_VGG on AZH dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.454	0.080	0.620	0.078	0.005	0.003	0.375	0.081
G: learned weights for RGB	0.377	0.083	0.542	0.094	0.018	0.024	0.479	0.102
RGB: adj. retinex	0.503	0.034	0.668	0.030	0.006	0.002	0.347	0.048
G: l. weights adj. retinex	0.304	0.103	0.456	0.126	0.006	0.004	0.581	0.137
RGB: gray world	0.409	0.066	0.577	0.067	0.004	0.002	0.444	0.070
G: l. weights gray world	0.415	0.061	0.584	0.061	0.011	0.005	0.458	0.120
G: Luma	0.116	0.050	0.203	0.082	0.014	0.019	0.815	0.095
G: Weights	0.179	0.077	0.296	0.106	0.017	0.014	0.737	0.099

Table 3 and Table 4 summarize the testing results of the LiteSeg_MobileNet model on the FUSeg dataset and AZH dataset, respectively. Similar to the UNet_VGG, the LiteSeg_MobileNet model also performs adequately with learned grayscale images compared to the RGB input.

6 Discussion

In this section, we discuss the effects of color on semantic segmentation in wound images. As described earlier, deep learning models often face challenges in the accurate segmentation of ulcer wound images due to limited data availability. Semantic segmentation focuses on pixel-level classification by assigning a label to each pixel in a given image. Therefore, local and global features are instrumental for accurate performance. Local features capture fine-grained details, helping the model to find wound contours or boundary edges, while global features focus on contextual information, understanding wound shapes and similar skin patterns.

We are investigating the impact of learned weights on a single-channel image, which could aid the model in identifying the most pertinent features early on. Figure 6 presents three images produced from an RGB image (see Fig. 5 (left)), using three different color conversion methods. Compared to the image on the left

Table 3. Performance comparison of LiteSeg_MobileNet on FUSeg dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.744	0.012	0.853	0.008	0.053	0.011	0.102	0.019
G: learned weights for RGB	0.675	0.025	0.806	0.018	0.069	0.016	0.144	0.041
RGB: adj. retinex	0.732	0.009	0.845	0.006	0.057	0.011	0.109	0.026
G: l. weights adj. retinex	0.676	0.027	0.806	0.019	0.085	0.020	0.125	0.041
RGB: gray world	0.733	0.013	0.846	0.009	0.059	0.013	0.104	0.020
G: l. weights gray world	0.674	0.031	0.805	0.022	0.077	0.022	0.136	0.030
G: Luma	0.229	0.106	0.361	0.135	0.023	0.006	0.673	0.115
G: Weights	0.259	0.083	0.405	0.108	0.052	0.057	0.597	0.107

Table 4. Performance comparison of LiteSeg_MobileNet on AZH dataset

Config	mIoU	Std IoU	mDice	Std Dice	mFPR	Std FPR	mFNR	Std FNR
RGB	0.369	0.071	0.535	0.075	0.005	0.003	0.493	0.077
G: learned weights for RGB	0.237	0.046	0.381	0.061	0.005	0.002	0.660	0.106
RGB: adj. retinex	0.440	0.037	0.610	0.035	0.007	0.004	0.415	0.059
G: l. weights adj. retinex	0.344	0.073	0.507	0.082	0.011	0.006	0.540	0.143
RGB: gray world	0.374	0.053	0.542	0.057	0.006	0.003	0.475	0.064
G: l. weights gray world	0.272	0.100	0.418	0.133	0.008	0.005	0.628	0.155
G: Luma	0.021	0.012	0.040	0.023	0.001	0.001	0.969	0.021
G: Weights	0.027	0.019	0.052	0.034	0.002	0.003	0.961	0.020

(BT-601 standard), the image in the middle (learned weighted image) amplifies the wound boundaries, which helps the model refine local features. Through this analysis, we conclude that machines can learn to interpret colors in many ways and differently than humans. This interpretation can be advantageous in improving the generalization properties of the trained ulcer segmentation model using color-to-grayscale conversion.

The quantitative results clearly show the difference in segmentation metrics between grayscale images obtained using traditional conversion methods and weights learned during the model training. For example, for the UNet_VGG model and FUSeg validation dataset (Table 1), the mean Dice value for grayscale images converted using learned weights is 0.845, while the same metric for traditional conversion methods is 0.536 and 0.614. Interestingly, the mean Dice value obtained for input RGB images was comparable to the result for grayscale images converted using learned weights (no statistical difference). The results obtained for the AZH test subset show lower values of all metrics. This could be understood since these images are much smaller and different, and examples from this domain were not used during training. However, for both models, much better results were obtained when adjusted retinex preprocessing was used (simple white balance) compared to the original RGB images. This demonstrates

the need for color calibration for RGB images (or huge training datasets with sufficient representation of all domains - often impossible) or the potential use of augmentation with grayscale images obtained for learned weights.

Furthermore, results from Tables 1, 2, 3, and 4 also show that fixed-weight methods are in line with human observation, while machines interpret the same color differently. The fixed-weight techniques do not help the deep learning models, leading to poor prediction performance. On the other hand, learned weights images present (in some configurations) comparable results with RGB images. Usually, the goal of research is to find the best model. In performed experiments, the UNet_VGG performed better for both datasets. For this model, the values of segmentation metrics for grayscale images obtained with learned weights were comparable with those obtained for RGB images. Therefore, the general conclusion is that such preprocessing is valuable and worth further investigation. The conversion weights can be learned for larger datasets and used in other preprocessing and augmentation experiments.

This study has several limitations, including a few datasets used in experiments, a limited number of models investigated, etc. We do not have many fixed-based conversion methods from three channels to a single channel, e.g., Cr, Cb, a^* , b^* , etc. However, preliminary results suggest the potential role of machine-oriented data preprocessing instead of traditional human-related color-to-grayscale conversion. Other methods can be compared in future research.

Other researchers used traditional color conversion methods to grayscale (e.g., [36–40]). In contrast, learned-based color-to-grayscale conversion methods can be potentially better for many tasks, mainly when fully explainable methods are applied for wound image analysis. In such cases, preprocessing using learned weights can be introduced instead of traditional conversion. Further, explainable image analysis methods can be used for wound shape analysis (e.g., monitoring of the healing process), wound classification, etc.

Thus, from these experiments, we validate potential applications of learned grayscale features in semantic segmentation, especially in ulcer wound segmentation, where wound shape, contour, edges, and texture attributes are very relevant. Additionally, learned grayscale images can partially reduce the complexity of the model, help the model learn the most critical features, and potentially avoid overfitting issues (they are less sensitive to color variations).

7 Conclusion

Early and precise assessment of ulcer wounds is crucial for proper treatment and management. Visual inspection by clinical professionals for diagnosis is labor-intensive and susceptible to human error. In contrast, computer-aided methods are highly efficient and can help early diagnosis. However, due to scarce data, deep learning methods face challenges of generalization and overfitting issues. This study aimed to address these challenges by exploring the impact of color on the segmentation of wound images. We designed a simple, universal model architecture to learn the task-specific weights for color-to-grayscale conversion.

We demonstrated that using learned weights for color-to-grayscale conversion leads to significantly better results in the semantic segmentation of wounds, achieving a mean Dice value of 0.845, compared to 0.536 and 0.614 for traditional RGB-to-grayscale conversion methods. For the best model, no statistical difference was found when comparing DICE metrics obtained for models trained on RGB images ($DICE = 0.868 \pm 0.048$) and grayscale images converted with learned weights (0.845 ± 0.020). The p-value calculated using the Mann-Whitney test for the FUSeg validation set was > 0.05 . We also demonstrated that color preprocessing using adjusted retinex can improve semantic segmentation results for similar but new domains for the applied (not tuned) model. Using grayscale images as input data presents competitive performance and lower computation complexity than RGB images. Additionally, using grayscale images as input data can potentially reduce the number of model parameters. The color-to-grayscale conversion weights can be learned for much more extensive (mono-domain or multi-domain) datasets and further used in other experiments for preprocessing, augmentation, and explainable image analysis methods.

References

1. Usman M (2025) Learned Weights Foot Ulcer Segmentation Repository. https://github.com/usmanraza121/Learned_weights_FUSeg
2. Sen CK (2019) Human wounds and its burden: an updated compendium of estimates. *Adv Wound Care* 8(2):39–48
3. Sun B et al (2022) An optimally designed engineering exosome-reductive COF integrated nanoagent for synergistically enhanced diabetic fester wound healing. *Small* 18(26):2200895
4. Boulton AJM (2008) The diabetic foot. In: *Controversies in treating diabetes: clinical and research aspects*, pp 251–267
5. Chang M, Nguyen TT (2021) Strategy for treatment of infected diabetic foot ulcers. *Acc Chem Res* 54(5):1080–1093
6. Armstrong DG et al (2023) Diabetic foot ulcers: a review. *Jama* 330(1):62–75
7. Patry J et al (2021) Outcomes and prognosis of diabetic foot ulcers treated by an interdisciplinary team in Canada. *Int Wound J* 18(2):134–146
8. Yan J et al (2022) Treatment of diabetic foot during the COVID-19 pandemic: a systematic review. *Medicine* 101(35):e30139
9. Oota SR et al (2021) Healtech-a system for predicting patient hospitalization risk and wound progression in old patients. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 2463–2472
10. Jørgensen LB et al (2016) Methods to assess area and volume of wounds—a systematic review. *Int Wound J* 13(4):540–553
11. Song B, Sacan A (2012) Automated wound identification system based on image segmentation and artificial neural networks. In: *2012 IEEE international conference on bioinformatics and biomedicine*, pp 1–4. <https://api.semanticscholar.org/CorpusID:47463>
12. Computerized segmentation and measurement of chronic wound images. *Comput Biol Med* 60:74–85. ISSN: 0010-4825. <https://doi.org/10.1016/j.compbiomed.2015.02.015>

13. Hu X et al (2019) Topology-preserving deep image segmentation. In: Advances in neural information processing systems, vol 32
14. Babu KS, Ravi KYB, Sabut S (2017) An improved watershed segmentation by flooding and pruning algorithm for assessment of diabetic wound healing. In: 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT). IEEE, pp 679–683
15. Wang C et al (2020) Fully automatic wound segmentation with deep convolutional neural networks. *Sci Rep* 10(1):1–9 Article number: 21897
16. Liu H et al (2022) Wound Segmentation with Dynamic Illumination Correction and Dual-view Semantic Fusion. *arXiv preprint* [arXiv:2207.05388](https://arxiv.org/abs/2207.05388)
17. Cao C et al (2023) Nested segmentation and multi-level classification of diabetic foot ulcer based on mask R-CNN. *Multimedia Tools Appl* 82(12):18887–18906
18. Ferreira Filipe et al (2021) A systematic investigation of models for color image processing in wound size estimation. *Computers* 10(4):43
19. Lam EY (2005) Combining gray world and retinex theory for automatic white balance in digital photography. In: Proceedings of the ninth international symposium on consumer electronics (ISCE 2005). IEEE, pp 134–139
20. Veredas FJ, Mesa H, Morente L (2015) Efficient detection of wound-bed and peripheral skin with statistical colour models. *Med Biol Eng Comput* 53:345–359
21. Loizou CP et al (2012) Evaluation of wound healing process based on texture analysis. In: 2012 IEEE 12th international conference on bioinformatics & bioengineering (BIBE). IEEE, pp 709–714
22. Song B, Sacan A (2012) Automated wound identification system based on image segmentation and artificial neural networks. In: 2012 IEEE international conference on bioinformatics and biomedicine. IEEE, pp 1–4
23. Fauzi MFA et al (2021) Segmentation and management of chronic wound images: a computer-based approach. In: Chronic wounds, wound dressings and wound healing, pp 115–134
24. Fauzi M et al (2015) Computerized segmentation and measurement of chronic wound images. *Comput Biol Med* 60:74–85
25. Foltynski P, Ciechanowska A, Ladyzynski P (2021) Wound surface area measurement methods. *Biocybern Biomed Eng* 41(4):1454–1465
26. Foltynski P, Ladyzynski P (2022) Digital planimetry with a new adaptive calibration procedure results in accurate and precise wound area measurement at curved surfaces. *J Diab Sci Technol* 16(1):128–136
27. Goyal M et al (2017) Fully convolutional networks for diabetic foot ulcer segmentation. In: 2017 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 618–623
28. Deng J et al (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
29. Hoiem D, Divvala SK, Hays JH (2009) Pascal VOC 2008 challenge. *World Lit Today* 24(1):1–4
30. Rania N, Douzi H, Yves L, Sylvie T (2020) Semantic segmentation of diabetic foot ulcer images: dealing with small dataset in DL approaches. In: El Moataz A, Mammass D, Mansouri A, Nouboud F (eds) *ICISP 2020*, vol 12119. LNCS. Springer, Cham, pp 162–169. https://doi.org/10.1007/978-3-030-51935-3_17
31. Cui C et al (2019) Diabetic wound segmentation using convolutional neural networks. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 1002–1005

32. Muñoz PL, Rodríguez R, Montalvo N (2020) Automatic segmentation of diabetic foot ulcer from mask region-based convolutional neural networks. *J Biomed Res Clin Invest* 1(1.1006) (2020)
33. Mahbod A et al (2022) Automatic foot ulcer segmentation using an ensemble of convolutional neural networks. In: 2022 26th international conference on pattern recognition (ICPR). IEEE, pp 4358–4364
34. Yi H et al (2022) OCRNet for diabetic foot ulcer segmentation combined with edge loss. In: *Diabetic foot ulcers grand challenge*. Springer, pp 31–39
35. Oota SR et al (2023) WSNet: towards an effective method for wound image segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 3234–3243
36. Cazzolato MT et al (2020) Semi-automatic ulcer segmentation and wound area measurement supporting telemedicine. In: 2020 IEEE 33rd international symposium on computer-based medical systems (CBMS). IEEE, pp 356–361
37. Garcia-Zapirain B et al (2017) Automated framework for accurate segmentation of pressure ulcer images. *Comput Biol Med* 90:137–145
38. Satheesha TY, Satyanarayana D, Prasad G (2015) Early detection of melanoma using color and shape geometry feature. *J Biomed Eng Med Imaging* 2(4):33
39. Bulan O, Artan Y (2016) Wheal detection from skin prick test images using normalized-cuts and region selection. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE, pp 1250–1253
40. Hettiarachchi NDJ et al (2013) Mobile based wound measurement. In: 2013 IEEE point-of-care healthcare technologies. IEEE, pp 298–301
41. Lukac R (2018) Single-sensor imaging: methods and applications for digital cameras. CRC Press
42. International Telecommunication Union (2011) Recommendation ITU-R BT.601-7: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. Technical report BT.601-7. ITU. https://www.itu.int/dms_pubrec/itu-r/rec/bt/r-rec-bt.601-7-201103-i!!pdf-e.pdf
43. International Telecommunication Unit (2015) Recommendation ITU-R BT.709-6
44. Sugawara M, Choi S-Y, Wood D (2014) Ultra-high-definition television (Rec. ITU-R BT.2020). *IEEE Signal Process Mag* 31(3):170–174
45. Wang C et al (2022) FUSeg: The Foot Ulcer Segmentation Challenge. arXiv preprint [arXiv:2201.00414](https://arxiv.org/abs/2201.00414) (2022)



Estimation of Body Ratios from Low-Resolution, Dual-Energy X-Ray Images Using Convolutional Neural Networks

Kamil Janczyk^{1,2(✉)}, Jacek Rumiński¹, Tomasz Neumann¹,
Aleksandra Szymczyk³, and Piotr Wiśniewski⁴

¹ FETI, Department of Biomedical Engineering, Gdansk University of Technology,
80-233 Gdansk, Poland
kamil.janczyk@pg.edu.pl

² Ship Hydromechanics Division, Maritime Advanced Research Centre, Gdansk,
Poland

³ Department of Emergency Medicine, Medical University of Gdansk, Gdansk,
Poland

⁴ Department of Endocrinology and Internal Diseases, Medical University of Gdansk,
Gdansk, Poland
<http://www.pg.edu.pl>

Abstract. Estimating body ratios from X-rays is an interesting research topic important for clinical applications such as assessing children's development or deviations from expected growth patterns. Research on these issues typically involves time-consuming and resource-intensive methods, such as manual body length measurements or high-resolution image annotations. We have proposed a preprocessing framework and adapted pre-trained convolutional neural network models to estimate the lengths of the body parts of children and young adults from low-resolution, dual-energy, whole-body, X-ray absorptiometry images. Furthermore, we proposed a solution for the automatic estimation of body ratios and the determination of samples significantly deviating from the average. The dataset of X-ray absorptiometry images was expanded with annotations of the length of individual body parts performed under the supervision of a specialist. The experimental results show that the proposed preprocessing techniques and the adapted convolutional neural network model achieved a mean relative error of up to 4.24 between the estimated and annotated lengths of body parts percent and the Lin correlation coefficient of convergence of 0.92. It was shown that, by employing low-resolution images and deep learning methods, one can accurately estimate the lengths of specific body parts, facilitating the precise detection of samples that deviate from the average.

Keywords: Dual-energy absorption · convolutional neural network · body ratios

1 Introduction

Dual-energy X-ray absorptiometry (DXA) is a medical imaging technique primarily used to assess bone health, but it has the potential to extend beyond this application [24]. In addition to quantifying bone mineral density, the technique allows the measurement of fat and lean body mass. It can generate low-resolution two-dimensional whole-body images, making it suitable for comprehensive analysis of body composition [31]. Anthropometric measurements, including height and segmental length, are fundamental tools for monitoring growth and development in children, and deviations from expected growth patterns can signal potential health problems, requiring further evaluation and intervention [6, 26]. Therefore, extracting anthropometric measurements from whole-body DXA images could provide additional valuable information on a child's physical development, nutritional status, and overall well-being.

This work is the following part of the larger project aimed at achieving the following research goal: the improvement of the acquisition of diagnostic knowledge from whole-body, low-resolution, dual-energy X-ray absorptiometry data, potentially related to a lower radiation dose for patients [16]. This knowledge could be extracted in many forms. For example, it could be a set of parameters: bone density (e.g. bone mineral density (BMD), bone mineral content (BMC)), fat density, muscle density, osteoporosis scores, bone age, etc.

2 Related Work

Many researchers focus on determining body lengths and ratios, as well as the correlation between them and health conditions. Most works on determining body ratios or lengths use manual body length measurements [11, 12] or image annotations [1, 7, 25]. The most commonly used ratios include the arm span divided by height (AS/H), height subtracted from arm span (AS/H), and the length of the upper segment divided by the length of the lower segment (U/L). The use of binary silhouettes for the automatic measurement of anthropometric dimensions was presented by Aslam et al. [4]. Dibra et al. [8] proposed the Human Shape Network model for the estimation of body shape from binary silhouettes or shaded images. The authors in [3] used machine vision systems with markers to determine the anthropometric lengths of body segments. Similarly, the application of a depth camera for automatic human body feature extraction is presented in [32]. In their work, Usha Chinappen-Horsley et al. [7] proposed two methods for measuring skeletal ratios (e.g., length of femur to height). They achieved inter- and intra-observer errors of 1. 6% for the linear pixel count method and 2. 3% for the reticule and ruler method. In [1], the authors present a study that describes strong linear correlations (Pearson correlation coefficient $r > 0.9$) between the five longitudinal measurements (body height and the lengths of four long bones: humerus, radius, femur, and tibia). They measured stature and limb using the 'ruler' tool of the DEXA software. Eszter Völgyi et al. [30] measured different widths and lengths of body segments

from DXA scan images using bones as references. Willem J.M. Gerver et al. [10] in their work used data from the Maastricht study to calculate the relative arm span together with height and assess body ratios to age and sex dependency (2,595 healthy Caucasian children, 52% boys, 48% girls, 2–17 years). As a result, they produced the graphs and reference tables of AS - H and AS/H for age and sex. The mean AS - H increased from -1.2 to $+1.5$ cm in boys and from -4.8 to $+1.6$ cm in girls. Mean AS/H increased from 0.9848 to 1.0155 in boys and from 0.9468 to 1.0028 in girls. Serap Turan et al. [29] measured height, arm span, upper and lower segments in 1302 healthy children (3–18 years). The mean values of the Us/Ls ratio decreased from 1.108 at three years to 0.922 at 15 years for boys. For girls, the US/LS ratio decreased to 0.946 at 13 years. Amitav Sarma et al. [22] investigated the relationship between arm span and standing height for males and females. In their work, they have shown that the standing height and arm span (measured manually) had a significant positive correlation in both males ($r = 0.988$) and females ($r = 0.991$). F. Yin and S. Zhou [33] proposed the Four-Stage Convolutional Neural Network that predicts body height and achieves an average accuracy equal to 99.1% for people in various positions and postures. Mosavi Sakina et al. [21] proposed an AdaBoost Regressor model for height estimation from 2D images, achieving a mean absolute error of 6.2 cm and a mean squared error of 7.9 cm. The authors [27] proposed a system for body parts measurement using the pose estimation method. Studies [2, 20] show a positive correlation coefficient between lower-limb lengths and stature.

2.1 Contribution

This work focused on body part ratio estimation for children and young adults from dual-energy X-ray images using a convolutional neural network. This could be important in many applications, including the proper assessment of a child's development. In the clinical evaluation of too tall or too short children, estimating body ratios may be helpful in assessing the likelihood of a primary growth disorder, such as skeletal dysplasia [22]. Moreover, research has shown a connection between anthropometric measurements, height, skeletal disproportion, and the increased risk of heart disease and diabetes [28]. The ratio of arm span to height (AS /H) is a potential indicator of age-related height loss and is directly correlated with the severity of shortness of breath [28]. Arm span measurements are an alternative in clinical scenarios where accurate standing height measurement poses challenges [22, 29]. Specific measurements, such as hip axis length, serve as potential predictors of fracture risk that are independent of bone mineral density [7, 30]. Our work presents a method for detecting patients with body ratios significantly deviating from the norm.

In this work, our contributions were as follows:

- A new annotated dataset was created, consisting of five annotated body parts (forearm, upper arm, distance between arms, torso, distance between pelvic joint and ankle joint).

- Two data preprocessing techniques are proposed for the original DXA, low-resolution data allowing pre-trained deep learning models to be adapted for the problem of body-ratio estimation.
- We proposed a method for automating the estimation of body ratios and determining samples significantly deviating from the average.
- We provided experimentally validated evidence that the ratios of parts of the body can be estimated from low-resolution DXA images using deep learning models. This can possibly make the process more automatic and provide interpretable information for the specialist.

The remainder of the paper is structured as follows. Section Methods describes the methodology used in the study. The results are presented and discussed in the Results and Discussion sections.

3 Methods

3.1 Data Acquisition

Data analyzed in this study were obtained using the Hologic QDR 4500 X-ray Bone Densitometer. It has a single switched-pulse, dual-energy X-ray tube that operates at 100 and 140kVp. The average tube current is 5 mA at 50% duty cycle. The maximum scan area is 195 cm $\times$ 65 cm - it allows scanning the lumbar spine (in anterior-posterior and lateral projections), the proximal femur, and the whole body. The original sample generated by the Hologic Densitometer contains high- and low-resolution data. In this study, we focus on low-resolution, whole-body data. Each sample contained six arrays of raw data. Each array corresponds to a not-preprocessed image of a high- or low-energy beam passing through a rotating drum filter and then a patient. The filter consisted of three different segments with radiopacities equivalent to those of tissue, bone, and air. The height of low-resolution images ranged from 84 to 128 pixels, and the width was constant - 106 pixels.

3.2 Dataset

The dataset consists of 1,455 whole-body DXA scan samples, but in accordance with the purpose of this study, we only used samples from young subjects. At the time of DXA scanning, 607 patients older than 230 months were omitted. In addition, 87 samples were removed because the patient's position made it impossible to label the scan correctly. Finally, the dataset includes samples 450 for males and 312 for females (the mean age was equal to 138.7 ± 42.3 months).

Data Preparation. The original DXA dataset consisted of raw data samples, each containing six data matrixes acquired using three filters for two energies of the beam: LOW_AIR, HIGH_AIR, LOW_BONE, HIGH_BONE, LOW_TISSUE, and HIGH_TISSUE. However, we decided to reduce the data

to three-channel images to apply pre-trained deep-learning models. The number of arrays was reduced by subtracting the arrays corresponding to high energy (HIGH) from arrays corresponding to low energy (LOW) for each filter. Next, each channel of the obtained images was normalized globally to a value from 0 to 255. As a result, RGB-like images were obtained, $R = (LOW\_AIR - HIGH\_AIR)$, $G = (LOW\_TISSUE - HIGH\_TISSUE)$, $B = (LOW\_BONE - HIGH\_BONE)$. Lastly, the images were symmetrically padded to 128×128 pixels. The padding was done with a constant value equal to the mean of 24 pixels selected from the background image separately for each channel.

During data preparation, we adopted three different approaches. First, the original 3-channel image with no additional preprocessing was used. In the second approach, the original image was transformed into a grayscale image. The grayscale intensity of each pixel was calculated using the following formula: $Intensity = 0.299 \times B + 0.587 \times G + 0.114 \times R$. To fit the model input, the grayscale image was tripled. In the third approach, we applied a contrast enhancement technique for each grayscale image. The Contrast Limited Adaptive Histogram Equalization (CLAHE) [5] was used. Based on the experimental results, we increased the brightness by a factor of 30 for this approach. The factor was selected from a range of tested values (10, 30, 50, 70). The best CLAHE parameters were selected after multiple experiments (clip limit was set to 7, and grid tile size was 8×8). Finally, in all approaches, the image was resized to 256×256 or 192×192 to fit the requirements for the input of the two adopted models. Bilinear interpolation was used to resize the images. Selected examples are presented in Fig. 1.

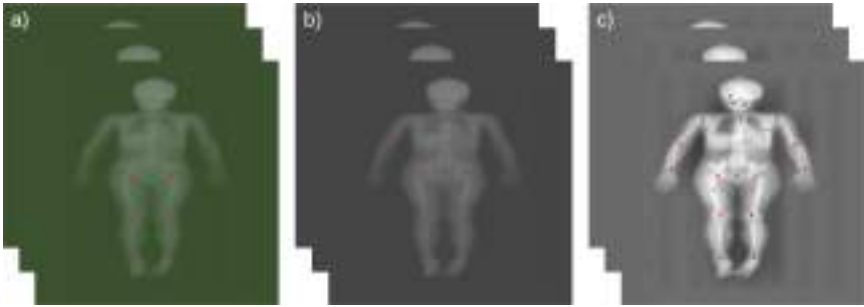


Fig. 1. Different variants of investigated images a) base image, b) grayscale image, c) grayscale + CLAHE image

Annotations. The annotations of the body parts carried out as part of this research endeavor were used to evaluate the effectiveness of the proposed framework. These annotations were performed under the supervision of a highly experienced medical specialist with 20 years of clinical practice. The whole dataset

was annotated twice by two different persons, who were instructed by the medical specialist. In the remaining part of the paper, we will refer to the labels as ‘label1’ and ‘label2’. Additionally, 10% of the data, randomly selected according to age and sex distribution, was annotated by an experienced medical specialist in this field (‘label3’). The length of the body’s individual parts was measured and labeled using ImageJ software[23]. The particular body parts were indexed, as shown in Fig. 2.

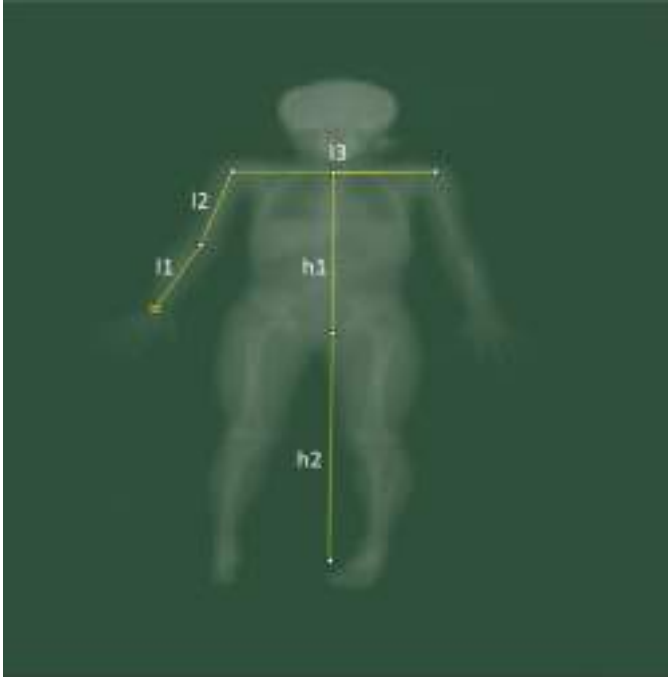


Fig. 2. Sample image with body length annotated. The labels correspond to the distance between individual joints: l1 - distance between the left wrist joint and the elbow joint, l2 - left the elbow joint and the shoulder joint, l3 - shoulder joints, h1 - the shoulder joints and the hip joints, h1 - the hip joints and the upper ankle joints.

3.3 Deep Learning Models

In this work, we adapted two neural network models: the Movenet Thunder [14,17] and the Movenet Lighting [15,17]. These are the convolutional neural network models that allow predicting human joint locations. Both models were pre-trained on the COCO (Common Objects in Context) dataset [19] and the Active dataset. The Active dataset consisted of images taken from YouTube fitness videos showcasing people exercising, stretching, and dancing. Based on joint

positions, it is possible to calculate the lengths of parts of the body corresponding to annotations.

Movenet consists of a MobileNetV2 image feature extractor with a Feature Pyramid Network decoder, followed by CenterNet prediction heads with custom post-processing logic. This model's input is a three-channel image with a size of 256×256 (Thunder) or 192×192 (Lightning).

The output consists of 17 key points (nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle) and the prediction confidence scores for each keypoint (in the range $[0.0, 1.0]$). Because some images have cropped heads, we skipped keypoints related to the head (nose, left eye, right eye, left ear, right ear).

The algorithm was implemented to measure segment lengths ($l1, l2, l3, h1, h2$) using the keypoints. The $l1, l2$, and $l3$ lengths were detected by calculating Euclidean distance between the corresponding joints: $l1$ length between left wrist and elbow, $l2$ length between left elbow and shoulder, and $l3$ length between left and right shoulder. Three additional midpoints were calculated to detect $h1$ and $h2$: the first between the shoulder joints, the second between the hip joints, and the third between the ankles. The $h1$ was calculated as the Euclidean distance between the former two midpoints, and the $h2$ as the distance between the latter two midpoints (Fig. 3). Euclidean distances were calculated using the NumPy library.

3.4 Metrics

The performance of each model and preprocessing approach was evaluated with the confidence score. The results of the estimation of the lengths of body parts using Movenet were confronted with annotations using the mean relative error (MRE) as a metric [13]. MRE was calculated according to Eq. 1.

$$MRE = \frac{x_l - x_m}{x_l} \cdot 100\% \quad (1)$$

where x_l - annotated length of the part of the body, x_m - estimated length of the part of the body.

The second metric used was Lin's Concordance Correlation Coefficient (R_{ccc}). R_{ccc} is a measure of how well a set of data compares to a "gold standard" measurement or test [18]. R_{ccc} was calculated according to the Eq. 2.

$$R_{ccc} = \frac{2rs_x s_y}{(\bar{x} - \bar{y})^2 + s_x^2 + s_y^2} \quad (2)$$

where r - Pearson correlation coefficient, s_x - variance of x , s_y - variance of y , $\bar{x}$ - mean of x , $\bar{y}$ - mean of y .

To better reflect the scale of differences between the annotation and estimation, we proposed the potential error of 1 pixel in preparation of the annotation ($1\text{PixelError} = \frac{1}{x_l} \cdot 100$). This error is defined as a MRE of 1 pixel of each annotated length ($l1, l2, l3, h1, h2$).

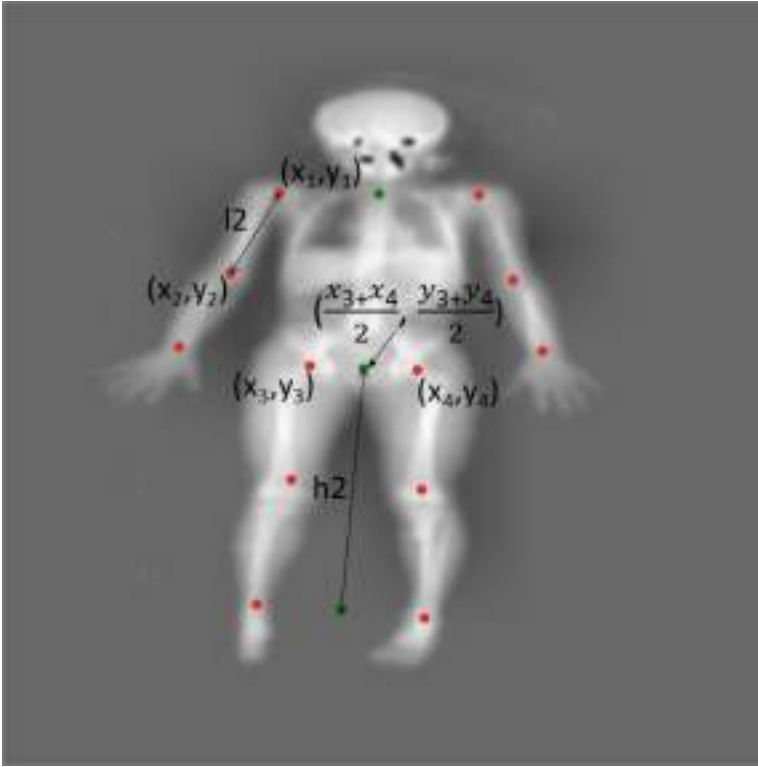


Fig. 3. Sample image with marked keypoints. The red dots correspond to keypoints from Movenet, where the green one is the midpoint between the hip and ankle joints. Detected lengths (e.g., $l2$ and $h2$) were calculated as Euclidean distances between pairs of points (e.g. (x_1, y_1) and (x_2, y_2)).

3.5 Testing Preprocessing and Models

The confidence score was obtained for each keypoint separately, and then the mean confidence score and the mean standard deviation were calculated. Without additional preprocessing, base images achieved a mean confidence score of 0.61 on the Movenet Thunder model. Converting the images to grayscale and applying the CLAHE improved the mean confidence score by 0.09 (up to 0.79) and reduced the mean standard deviation. Using the Movenet Lightning model generated slightly worse results for the same preprocessing approach (confidence score equal to 0.75) - Table 1.

3.6 Comparison with Annotations

Using the best estimation results (grayscale + CLAHE, model Thunder) and all labels, the mean relative error was calculated for each determined length (Table 2). Additionally, the 1-pixel error value was calculated (only for 'label3').

Table 1. Mean confidence score with mean with standard deviation achieved for different variants and models

Confidence score	
base - Thunder	0.61 ± 0.16
grayscale - Thunder	0.70 ± 0.14
grayscale + CLAHE - Thunder	0.79 ± 0.10
grayscale + CLAHE - Lightning	0.75 ± 0.11

Table 2. MRE [%] with standard deviation for different labels and the Movenet (model Thunder with grayscale + CLAHE preprocessing approach)

	MRE for Movenet			1 pixel error
	'label1'	'label2'	'label3'	
l1	9.75 ± 5.51	15.39 ± 6.94	8.37 ± 5.64	5.97 ± 0.94
l2	6.87 ± 5.35	10.26 ± 6.54	7.27 ± 5.32	4.77 ± 0.80
l3	8.41 ± 4.18	8.43 ± 6.07	4.24 ± 3.43	2.42 ± 0.34
h1	8.16 ± 6.85	10.71 ± 7.89	13.81 ± 6.94	2.72 ± 0.40
h2	12.89 ± 7.0	5.31 ± 4.03	5.10 ± 4.63	1.97 ± 0.33

Comparing the Movenet results to different labels, the MRE ranged from $4.24\% \pm 3.43\%$ for h2, 'label3', to $15.39\% \pm 6.94\%$ for l1, 'label2'. The maximum and minimum 1-pixel error equals $5.97\% \pm 0.94\%$ and $1.97\% \pm 0.33\%$, respectively. For 'label2' and 'label3', it can be noticed that h1 is an exception in the trend that the longer lengths tend to have greater MRE. Using 'label3' as ground truth, the mean relative error was calculated for 'label1' ($10.11\% \pm 4.00\%$) and for 'label2' ($8.14\% \pm 5.55\%$).

Lin's Concordance Correlation Coefficient was calculated between 'label3' and the rest of the sets (Table 3).

Table 3. R_{ccc} for 'label3' and rest sets for each body part length.

	'label3'					
	l1	l2	l3	h1	h2	
'label1'	0.88	0.88	0.76	0.62	0.63	
'label2'	0.71	0.80	0.77	0.92	0.95	
Movenet	0.75	0.83	0.92	0.41	0.90	

For Movenet $R_{ccc} \geq 0.90$ was obtained for two body part lengths, l3 (0.92 - Fig. 4) and h2 (0.90 - Fig. 5). For 'label2', this was achieved for h1 (0.92) and h2 (0.95). R_{ccc} exhibited values lower than 0.9 across all lengths for 'label1'.

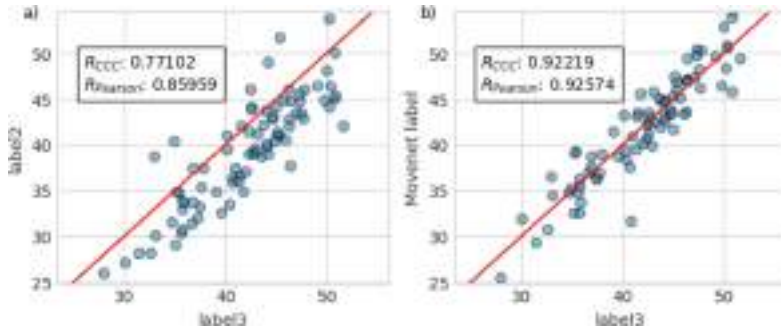


Fig. 4. Samples distribution for length l3: a) ‘label2’ vs ‘label3’, b) Movenet vs ‘label3’

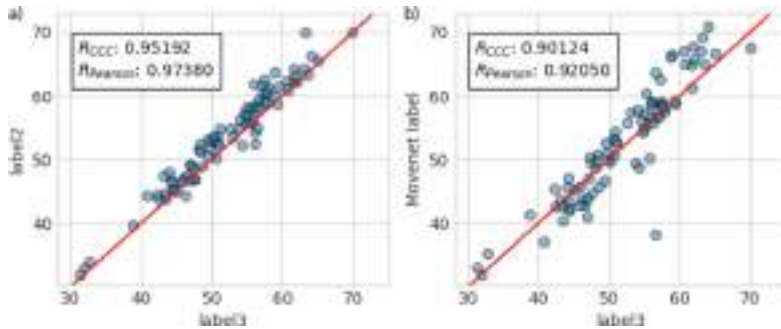


Fig. 5. Samples distribution for length h2: a) ‘label2’ vs ‘label3’, b) Movenet vs ‘label3’

3.7 Analysis of Outliers

Based on the obtained lengths, the arm span ($AS = l1*2 + l2*2 + l3$) and height ($H = h1 + h2$) were calculated, and the dependence between body ratios and age were plotted for both labels and the Movenet. The ratios used included AS/H , $AS-H$ and U/L ($U = h1$, $Ls = h2$). Lists of samples that deviated by more than three standard deviations from the 48-month moving average (outliers) were made (an example is shown in Fig. 6).

The number of outliers for labels (‘label1’ and ‘label2’) and Movenet, together with the number of the same samples occurring in both lists of outliers, is shown in Table 4, separately for males and females. The Merged list was also created by combining outliers from 3 basic ratios and removing duplicates for both labels and the Movenet.

The Movenet identified the highest number of repeated outlier samples for the AS/H ratio. For men, the Movenet correctly identified 5 out of 10 and 4 out of 7 repeated outliers for ‘label1’ and ‘label2’, respectively. For women, the corresponding numbers were 1 out of 3 and 1 out of 2. In the merged list, the number of repeated outliers was 8 out of 17 and 7 out of 14 for men, and 2 out of 6 and 2 out of 5 for women, for ‘label1’ and ‘label2’, respectively.

Table 4. The numbers of outliers were defined as follows: L - number of outliers for the label; M - number of outliers for Movenet; R - number of same outliers for the Movenet and label; M1 - males, ‘label1’; F2 - females, ‘label2’, etc.

	AS /H		AS - H		US/LS		Merged	
	L	MR	L	MR	L	MR	L	MR
M1	10	9	5	6	5	1	7	8
M2	7	9	4	6	5	2	7	8
F1	3	3	1	2	3	0	3	2
F2	2	3	1	2	3	1	2	2

Notably, the number of outliers for women is disproportionately smaller, even when considering their lower representation in the dataset. In the borderline case - AS/H for ‘label1’, the number of outliers for women is 3, while for men it is 10.

4 Discussion

The proposed solution achieved satisfactory results for the estimation of body ratios (MRE ranging from $15.39\% \pm 6.94\%$ to $4.24\% \pm 3.43\%$) and detection of samples significantly deviating from the average (up to 4 out of 7 repeating outliers) using the whole-body, low-resolution DXA image. First, the MRE and standard deviation achieved for the Movenet is comparable to the difference between annotations made by a specialist (‘label3’) and a non-specialist (‘label1’ - $10.11\% \pm 4.00\%$ and ‘label2’ - $8.14\% \pm 5.55\%$). Second, the value of the 1-pixel error (ranging from 5.97% to 1.97%) shows that the potential error in the annotation can greatly contribute to the final evaluation of the model. Based on the 1-pixel error, the error of the adopted model in estimating the lengths of the body parts was approximately 2 to 4 pixels (with the exception of h2 for ‘label1’ (<6 pixels) and h1 for ‘label3’ (<5 pixels)). The Movenet estimations and the ‘label1’ and ‘label2’ sets were compared with label3 using Lin’s Concordance Correlation Coefficient. When comparing the R_{ccc} calculated for ‘label1’ and ‘label2’, it can be seen that the annotations ‘label2’ agree better with ‘label3’. The difference can be especially noticeable for the h1 - 0.62 vs 0.92 and the h2 - 0.63 vs 0.95 for ‘label1’ and ‘label2’ accordingly. For the l1, l2, and h2 lengths, the R_{ccc} achieved for the Movenet did not deviate from those obtained by non-specialists. For the l3, the R_{ccc} was equal to 0.92 for the Movenet, 0.76 for ‘label1’, and 0.77 for ‘label2’. For h1, the Movenet achieved the lowest R_{ccc} (0.41). Due to the differences between R_{ccc} for h1 and the other lengths, three samples with the highest MRE for h1 were examined individually. At first, they were sent to specialists to annotate key points. Then, by comparing the position of the key points annotated by the specialist with those estimated by the Movenet, we determined that the neural network places the hip joint key points too high, leading to a shortening of the length h1 (Fig. 7).

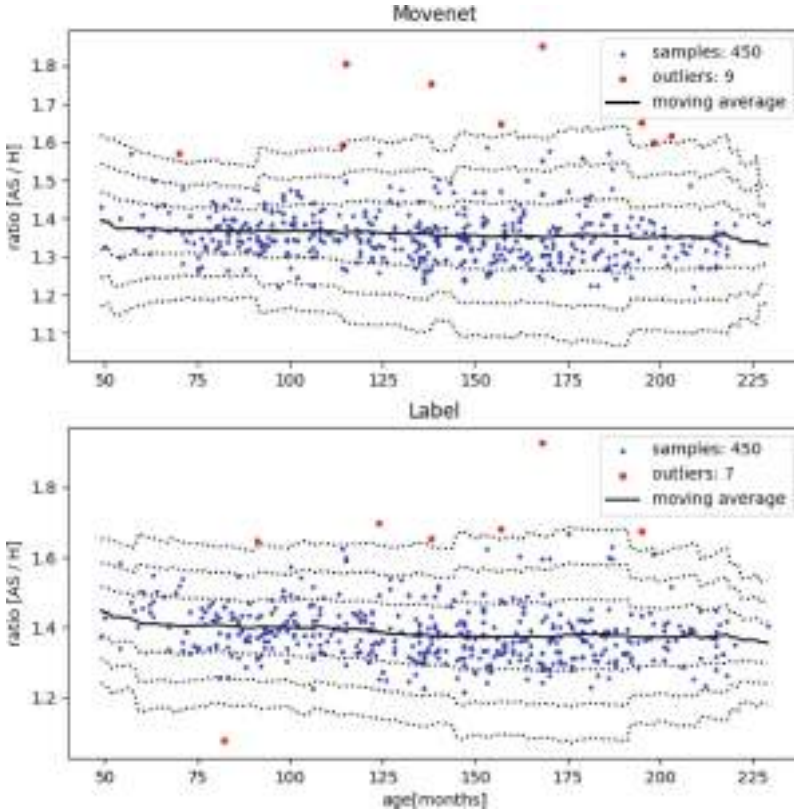


Fig. 6. The plot of moving average of AS/H body ratio vs. age along with standard deviation lines - 'label2', males.

In this work, we employed three distinct preprocessing approaches: using the original image, converting to grayscale, and applying grayscale conversion, followed by contrast and brightness enhancement. The optimal results were obtained with the third approach, where images were converted to grayscale, followed by contrast enhancement using the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique and a brightness increase with a selected factor ($K=30$). The baseline results for the original images were 0.61 ± 0.16 . After converting the images to grayscale, the confidence score improved to 0.70 ± 0.14 , and with the additional application of the CLAHE technique, the score further increased to 0.79 ± 0.10 . The parameters for CLAHE and the brightness factor were experimentally determined based on the confidence score. The preprocessing methods are considered optimal for a given experiment configuration based on the obtained confidence score values. Additionally, two models were evaluated, the Movenet Thunder [14] and the Movenet Lightning [15]. The former one achieved slightly better results (confidence score equal to 0.79). Finally, we

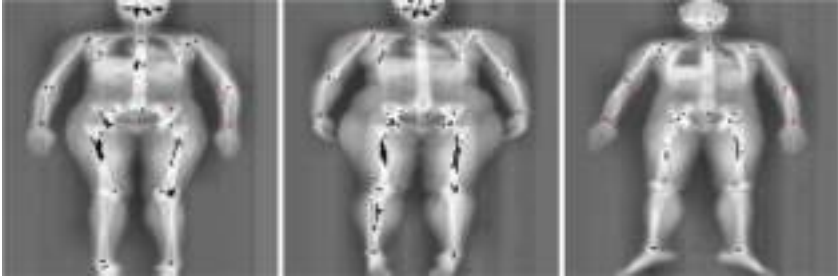


Fig. 7. Three samples with the highest MRE for h1 length. The red dots correspond to keypoints from the Movenet, the green one to the midpoint, and the blue one to the annotation from the specialist.

investigated three different body ratios (AS/H, AS - H, and US/LS). For each, the average ratio-age curve was plotted together with the standard deviation for ‘label1’, ‘label2’, and the Movenet estimations. Then, the repeatability of the outliers obtained for the Movenet and both sets of labels were compared. The best repeatability was achieved for the AS/H body ratio. For men, Movenet correctly indicated 5 out of 10 outliers (‘label1’) and 4 out of 7 (‘label2’) outliers (‘label2’). For women, it was 1 out of 3 and 1 out of 2, for ‘label1’ and ‘label2’ respectively).

Previous literature on determining body length and ratios featured manual measurements (e.g., [10,29]) or image annotations(e.g., [1,30]). In [7] for measuring skeletal ratios, the authors achieved inter- and intra-observer errors of 1.6% for the linear pixel count method and 2.3% for the reticule and ruler method. However, it is important to indicate that in our work, we proposed an automated solution that achieved results comparable to those of a non-specialist in determining body lengths. It is also important to underline the low-resolution images used in our work, with heights ranging from 84 to 128 pixels and constant widths equal to 106 pixels.

This study has several limitations. Some of them concern the resolution of the images used. In the future, the results for low-resolution and high-resolution images for selected body parts will be compared (owing to the content of high-resolution images). However, the application of low-resolution images has many advantages, such as lower doses, less energy, faster processing, etc. The following work might also validate the application of super-resolution methods [9] and investigate their roles in the accuracy of measurements of body parts from dual X-ray images. Additionally, in this work, we used only pre-trained models. Perhaps the fine-tuning could improve results for estimating the length of body parts from low-resolution images, considering the results obtained in reference to one-pixel accuracy.

4.1 Conclusion

It was shown that it was possible to estimate the body ratios based on their low-resolution image obtained from a DXA scan using the Movenet model. The outliers of body ratios detected using the Movenet model correspond to outliers from annotations made by hand.

The preliminary studies indicate that preprocessing methods-grayscale conversion, histogram equalization, and brightness enhancement-enhance estimation accuracy compared to the original, non-preprocessed DXA images. However, for images obtained with devices from other manufacturers, different settings and calibration procedures may be in use (possible domain change). So, preprocessing methods should be further developed to ensure consistency and adaptability across various imaging systems. This is the standard procedure for all types of medical imaging methods. In the future, deep learning-based domain adaptation methods can overcome this limitation.

Future work on the subject of body ratios should focus on the analysis of outliers to determine conditions causing them to deviate from the average. Then, additional experiments should be conducted using higher-resolution images. As for the whole project concerning improving the acquisition of diagnostic knowledge from dual-energy X-ray absorptiometry data, we plan to focus on predicting the different parameters (e.g., bone mineral density) coupled with image segmentation.

References

1. Abrahamyan DO, Gazarian A, Braillon PM (2008) Estimation of stature and length of limb segments in children and adolescents from whole-body dual-energy X-ray absorptiometry scans. *Pediatr Radiol* 38(3):311–315. <https://doi.org/10.1007/s00247-007-0720-x>
2. Ahmed AA (2013) Estimation of stature using lower limb measurements in Sudanese Arabs. *J Forensic Leg Med* 20(5):483–488. <https://doi.org/10.1016/j.jflm.2013.03.019>
3. Arellano-González JC, Medellín-Castillo HI, Hernández-Molinar RI, Plascencia-Mora H, González-de AA (2023) Determination of anthropometric lengths of body segments using machine vision systems. *Machines* 11:369. <https://doi.org/10.3390/machines11030369>
4. Aslam M, Rajbdad F, Khattak S, Azmat S (2017) Automatic measurement of anthropometric dimensions using frontal and lateral silhouettes. *IET Comput Vis* 11:434–447. <https://doi.org/10.1049/iet-cvi.2016.0406>
5. Bradski G (2000) The OpenCV library. *Dr Dobb's J Softw Tools Prof Program* 25(11):120–125
6. Casadei K, Kiel J (2022) Anthropometric measurement. StatPearls Publishing, Treasure Island. PMID: 30726000
7. Chinappen-Horsley U, Blake GM, Fogelman I, Spector TD (2007) A method for determining skeletal lengths from DXA images. *BMC Musculoskelet Disord* 16; 8(1). <https://doi.org/10.1186/1471-2474-8-113>

8. Dibra E, Jain H, Öztireli C, Ziegler R, Gross M (2016) HS-Nets: estimating human body shape from silhouettes with convolutional neural networks. In: 2016 fourth international conference on 3D vision (3DV), 25–28 October 2016, Stanford, CA, USA, pp 108–117. <https://doi.org/10.1109/3DV.2016.19>
9. Dong C, Loy CC, He K, Tang X (2016) Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell* 38(2):295–307. <https://doi.org/10.1109/tpami.2015.2439281>
10. Gerver W, Gkourogianni A, Dauber A, Nilsson O, Wit JM (2020) Arm span and its relation to height in a 2- to 17-year-old reference population and heterozygous carriers of ACAN variants. *Horm Res Paediatr* 93(3):164–172. <https://doi.org/10.1159/000508500>
11. Gerver WJM (1988) Measurement of the body proportions in children (the Oost-erwolde study), 187 p
12. Gerver W, de Bruin R (2003) Growth velocity: a presentation of reference values in Dutch children. *Horm Res Paediatr* 60(4):181–184. <https://doi.org/10.1159/000073230>
13. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D et al (2020) Array programming with NumPy. *Nature* 585(7825):357–362. <https://doi.org/10.1038/s41586-020-2649-2>
14. Internet. Movenet Thunder. <https://tfhub.dev/google/movenet/singlepose/thunder/4>. Accessed 8 Jan 2025
15. Internet. Movenet Lightning. <https://tfhub.dev/google/movenet/singlepose/lightning/4>. Accessed 8 Jan 2025
16. Janczyk K, Rumiński J, Neumann T, Głowacka N, Wiśniewski P (2022) Age prediction from low resolution, dual-energy X-ray images using convolutional neural networks. *Appl Sci* 12(13):6608. <https://doi.org/10.3390/app12136608>
17. Joshi RB, Joshi D (2021) MoveNet: a deep neural network for joint profile prediction across variable walking speeds and slopes. *IEEE Trans Instrum Meas* 70:1–11. <https://doi.org/10.1109/tim.2021.3073720>
18. Lin LI-K (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45(1):255. <https://doi.org/10.2307/2532051>
19. Lin T-Y et al (2014) Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) ECCV 2014, vol 8693. LNCS. Springer, Cham, pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
20. Özaslan A, İscan MY, Özaslan İ, Tuğcu H, Koç S (2003) Estimation of stature from body parts. *Forensic Sci Int* 132(1):40–45. [https://doi.org/10.1016/s0379-0738\(02\)00425-5](https://doi.org/10.1016/s0379-0738(02)00425-5)
21. Sakina M, Muhammad I, Abdullahi SS (2024) A multi-factor approach for height estimation of an individual using 2D image. *Procedia Comput Sci* 231:765–770. <https://doi.org/10.1016/j.procs.2023.12.140>
22. Sarma A, Barman B, Das G, Saikia H, Momin A (2020) Correlation between the arm-span and the standing height among males and females of the Khasi tribal population of Meghalaya state of north-eastern India. *J Family Med Prim Care* 9(12):6125. https://doi.org/10.4103/jfmpc.jfmpc_1350_20
23. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH image to imagej: 25 years of image analysis. *Nat Methods* 9(7):671–675. <https://doi.org/10.1038/nmeth.2089>
24. Shepherd J, Crabtree NJ (2016) Dual-energy X-ray absorptiometry technology. In: Fung EB, Bachrach LK, Sawyer AJ (eds) *Bone Health Assessment in Pediatrics*. Springer, Cham, pp 53–74. https://doi.org/10.1007/978-3-319-30412-0_3

25. Sievänen H, Kannus P, Oja P, Vuori I (1994) Dual energy X-ray absorptiometry is also an accurate and precise method to measure the dimensions of human long bones. *Calcif Tissue Int* 54(2):101–105. <https://doi.org/10.1007/bf00296059>
26. Singh SP, Mehta P (2009) Human body measurements: concepts and applications. PHI Learning, New Delhi
27. Singh D, Panthri S, Venkateshwari P (2022) Human body parts measurement using human pose estimation. In: 2022 9th international conference on computing for sustainable global development (INDIACom). <https://doi.org/10.23919/indiacom54597.2022.9763292>
28. Tan MP, Bansal SK (2012) The arm span to height relationship and its health implications. In: *Handbook of Anthropometry*, pp 741–754. https://doi.org/10.1007/978-1-4419-1788-1_41
29. Turan S, Bereket A, Omar A, Berber M, Ozen A, Bekiroglu N (2005) Upper segment/lower segment ratio and armspan-height difference in healthy Turkish children. *Acta Paediatr* 94(4):407–413. <https://doi.org/10.1111/j.1651-2227.2005.tb01909.x>
30. Völgyi E, Tylavsky FA, Xu L, Lu J, Wang Q, Alén M et al (2010) Bone and body segment lengthening and widening: a 7-year follow-up study in pubertal girls. *Bone* 47(4):773–782. <https://doi.org/10.1016/j.bone.2010.07.007>
31. Wells JC (2005) Measuring body composition. *Arch Dis Child* 91(7):612–617. <https://doi.org/10.1136/ad.2005.085522>
32. Xiaohui T, Xiaoyu P, Liwen L, Xia Q (2018) Automatic human body feature extraction and personal size measurement. *J Vis Lang Comput* 47. <https://doi.org/10.1016/j.jvlc.2018.05.002>
33. Yin F, Zhou S (2020) Accurate estimation of body height from a single depth image via a four-stage developing network. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). <https://doi.org/10.1109/cvpr42600.2020.00829>



Artificial Intelligence Models for Age and Sex Estimation from Electrocardiograms

Daniel Francisco Alborno¹ , Cristian Ariel Sposato¹ ,
Gustavo Javier Meschino¹ , Héctor Santiago Manzollilo² ,
and Virginia Laura Ballarín¹

¹ FASTA University, Mar del Plata, Argentina
{gmeschino, vballari}@ufasta.edu.ar

² Cardiology Institute of Corrientes, Corrientes, Argentina

Abstract. The electrocardiogram (ECG) is a fundamental tool in cardiovascular evaluation, widely used in clinical settings due to its accessibility, low cost, and ease of application. Recent advances in artificial intelligence (AI) have expanded the potential applications, enabling the estimation of demographic characteristics such as age and sex from ECG signals. This study aims to develop and validate an AI model capable of estimating age and sex from ECG records in a local population, evaluating its performance compared to previous studies and analyzing its potential impact on clinical applications. A convolutional neural network (CNN) was trained using a dataset of 275,000 anonymized ECGs. The model achieved an 87% accuracy in sex classification and an average absolute error of 9 years for age estimation. These results suggest that the predicted age may serve as a biomarker for biological aging and cardiovascular risk. The integration of this AI tool into clinical practice could lead to more efficient patient stratification and optimize healthcare resource allocation.

Keywords: Electrocardiogram · Artificial Intelligence · Age Estimation · Sex Classification · Machine Learning

1 Introduction

The electrocardiogram (ECG) is a widely used, non-invasive tool in cardiovascular medicine, found in both small and large populations throughout the country. It is a non-invasive, painless test that can be performed using low-cost electrocardiograph devices and supplies [1]. Conducting an electrocardiogram requires minimal technical training and can be performed by healthcare personnel or trained technicians. It reflects the structural and functional state of the heart, which cannot be altered by the user. The electrocardiogram is reproducible and remains highly consistent in the same patient unless the heart undergoes changes due to disease. In essence, it provides a direct reflection of cardiac activity, with all visible and hidden data, without intermediaries between the cardiac cells and the recorded electrocardiographic paper. These characteristics make it a valuable study for computer-based analysis when transformed into individual data and later into databases of multiple patients. These data are of high quality as they undergo minimal adaptations from acquisition to their aggregation in indexed datasets.

Age and sex are fundamental pieces of information in medical records, often serving as key variables in the diagnostic process. In medical records, they are among the first variables to be recorded. Given the vast number of possible diseases a patient may present, diagnosis is carried out in stages, usually leading to two or three possible diagnoses. This complex process involves categorizing the individual into groups based on data obtained primarily through in-person interactions and interviews with a physician. Each group has a certain probability of presenting typical diseases (statistically more likely), which guides the diagnostic process used in medical consultations. Age and sex are among the most powerful variables for differentiating potential diagnoses.

Artificial intelligence (AI), when properly trained and used as a tool, is powerful, fast, efficient, precise, and scalable. Today, it is widely accessible via the internet or smart devices, making it ubiquitous. AI can rapidly interpret a vast number of medical tests with minimal supervision and low-cost resources.

Currently, electrocardiograms are not used to determine a person's age or sex because this information is hidden from physicians' direct observation. However, this could change with a well-designed and well-trained AI model. But why use an electrocardiogram and AI to answer a question that could be resolved in seconds by simply asking the patient?

Recently published studies by Mayo Clinic in the United States (and others) gained attention by demonstrating that this is possible with remarkable accuracy [2–5]. The researchers also found that when the model's predicted age did not correlate with the patient's actual chronological age, the patient had a different life expectancy than expected. In other words, they identified a discrepancy between the patient's chronological age and their estimated age according to the model. This calculated age could be considered as a "biological age," as patients whose calculated age was higher than their chronological age tended to be significantly sicker. Essentially, the model was not incorrect but rather reflected a different age, the biological age. It is common to see individuals whose chronological age does not match an observer's subjective estimate, leading to remarks like, "They look younger." This occurs because aging progresses at different rates for different individuals. Until now, these differences were subjective, but an AI model could provide an objective measure, and according to the mentioned study, this has prognostic implications.

If these findings are validated in different populations, such a model could, for example, help prioritize patients who need earlier, faster, and more thorough medical supervision. It could serve as an assistant for healthcare professionals responsible for treating large numbers of patients, improving efficiency in public health.

One of the most important aspects of scientific research is the reproducibility of experiments, while AI models require validation [6]. The recently published model was trained on a specific population from a distant country's medical center.

Public health operates with finite resources. A model capable of automatically analyzing large numbers of electrocardiograms and identifying individuals whose biological age exceeds their chronological age could save many hours of highly specialized work. A recent estimate suggests that AI-based technologies could reduce healthcare professionals' workload by up to 30% [7]. This work aims to determine whether an AI model can assist healthcare professionals in identifying cases requiring greater attention due to

poor prognosis in the local population, using local data. Validating this finding would reduce biases and enhance the model's reliability.

2 Materials and Methods

2.1 Study Design

This retrospective study utilized anonymized electrocardiographic (ECG) records obtained from the *Instituto de Cardiología de Corrientes*. Figure 1 shows a typical recording study, highlighting the spatial distribution of the 12 leads using rectangular boundaries.



Fig. 1. Typical image of a ECG study and regions of the leads.

The study design focused on developing an AI model capable of estimating patient age and sex directly from ECG recordings stored in digital image format, within PDF files.

The original PDF documents included one page of size of 247.5×191.25 mm, within which a non-embedded binary image of 3300×2550 pixels was contained. This configuration yielded an approximate resolution of 339 DPI. For subsequent analysis, the binary image was exported from the PDF at a resolution of 300 DPI. This export was chosen to provide a consistent and slightly lower resolution for further processing, balancing detail retention with computational efficiency.

Given the image-based format of the data, several approaches were considered and evaluated for feature extraction and model development:

- **Whole-Image Feature Extraction:** This approach treated the entire ECG image as a single entity, trying to discover global patterns and morphological features that could correlate with age and sex. This method aimed to capture holistic visual characteristics of the ECG. For whole-image feature extraction, techniques from convolutional processing were explored.

- **Individual Lead Feature Extraction:** Each of the twelve ECG leads (derivations) was processed separately as an individual image. This allowed the model to focus on lead-specific characteristics, potentially revealing subtle variations in electrical activity related to age and sex that might be masked in the whole-image approach. This approach explored the hypothesis that analyzing leads individually could improve discriminative power. For individual lead feature extraction, similar analysis as the previous option was applied to each lead separately and concatenating the features
- **Signal Extraction and Digital Signal Processing (DSP):** This approach involved digitally reconstructing the ECG signals from the image files. This enabled the application of traditional signal processing techniques and time-series analysis to extract features in the time domain. This method explored the potential of considering established DSP methods for ECG analysis.

The ECG data was available in anonymized PDF format due to previous data storage practices. An interdisciplinary team comprising cardiologists, engineers, and AI researchers collaborated to design, develop, and validate the model. This collaborative approach ensured clinical relevance and facilitated the evaluation of the different feature-extraction strategies.

Based on preliminary experiments, the Signal Extraction and Digital Signal Processing (DSP) approach demonstrated superior performance in both age and sex estimation and was therefore selected for further development. This approach allowed us to focus on established time-series analysis techniques and extract features from the ECG signals. Extracted lead signals from the recordings are shown in Fig. 2.



Fig. 2. Extracted 12 lead signals.

The comparative evaluation focused on identifying the most effective strategy for age and sex estimation while considering computational efficiency and practical applicability in clinical settings. While the whole-image and individual lead approaches were explored, the DSP method provided a richer set of features and ultimately yielded the best performance.

The following section details the development and training of the CNN model based on the DSP approach for feature extraction.

2.2 Data Collection

This retrospective study utilized anonymized electrocardiographic records obtained from the *Instituto de Cardiología de Corrientes*. The dataset comprised 275,000 ECGs stored in PDF format. These PDF files contained images of the 12-lead ECG recordings.

The study included ECGs from patients aged 18 years and older. This age restriction was implemented to focus on adult ECG patterns and exclude pediatric ECGs, which can have distinct characteristics.

Several preprocessing steps were applied to the ECG data to prepare it for model training:

- **Anonymization:** To ensure patient privacy and comply with ethical guidelines, all ECG records underwent a rigorous anonymization process before analysis. This process, carried out by the Institute operator, involved the removal of all identifying patient information from the ECG files. Each record was then assigned a unique identifier, decoupling the data from any personal identifier. The authors or developers did not have access to the original, nonanonymized data, ensuring the confidentiality and protection of patient information throughout the study.
- **Signal Extraction:** Since the ECGs were in PDF format, the first step involved digitally reconstructing the 12-lead ECG signals from the image files. This was achieved using a custom algorithm based on image processing techniques.
- **Artifact Removal:** Artifacts were removed using filtering techniques and assuring the continuity of the signals. Cases where recordings of the leads were out of the limits (due to recording problems or other cases) were deleted from the database.
- **Data Quality Control:** A cardiologist reviewed a subset of the ECGs to ensure the accuracy of the signal extraction and the effectiveness of the artifact removal process. ECGs with excessively noisy or distorted signals that could not be adequately cleaned were excluded from the study.

After the automated processing steps, records affected by acquisition or processing anomalies were removed to ensure data integrity. From the remaining dataset, 119,845 records were kept, considering preservation of a balanced distribution across sex and age.

2.3 Model Development

A Convolutional Neural Network (CNN) architecture was chosen for the analysis of the ECG signals [8]. CNNs are well-suited for processing data with grid-like structures, such as images or time series, and have demonstrated strong performance in various medical image analysis tasks [9]. While initial explorations considered Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) commonly used for sequential data, computational constraints and the availability of resources led to the selection of a CNN for its efficiency and established effectiveness in feature extraction from ECG-like data.

The training process involved optimizing the model's parameters to minimize the difference between the predicted and actual age and sex. Various hyperparameters, such as the learning rate, batch size, and number of epochs, were tuned to achieve optimal performance. The training progress was monitored using metrics such as accuracy for sex classification and mean absolute error (MAE) for age estimation.

Multiple CNN architectures were explored by varying the number of convolutional layers, filter sizes, number of filters, and the configuration of the fully connected layers. The performance of each model was evaluated on the validation set, and the model with the lowest MAE for age estimation and highest accuracy for sex classification was selected as the final model.

The final CNN model architecture consisted of the following components:

- **Temporal Filtering:** Six stages of temporal filtering were implemented using 1-dimensional convolutional filters. These filters were designed to capture local temporal patterns and features within each ECG lead. The filter lengths, denoted as K , were [7, 5, 5, 5, 3, 3], and the corresponding number of filters, denoted as N , were [16, 16, 32, 32, 64, 64]. Each convolutional layer was followed by a max pooling operation to down sample the feature maps and reduce the dimensionality of the data.
- **Spatial Filtering:** A spatial filtering stage was incorporated using 128 convolutional filters. This stage aimed to combine information across the different ECG leads, capturing relationships and dependencies between them.
- **Fully Connected Layers:** Two fully connected layers, also known as dense layers, were added to the network. These layers processed the extracted features and learned higher-level representations. The final layer consisted of a single neuron with a linear activation function for age estimation and a sigmoid activation function for sex classification.

The model was trained on preprocessed electrocardiogram (ECG) data, partitioned into training (80%) and testing (20%) sets. The demographic distribution of the dataset, including sex and age distribution by decade, is presented in Fig. 3. To monitor generalization performance during training, a randomly selected 15% subset of the training set was reserved as a validation set.

The model development process faced several challenges, including computational limitations, data quality issues, and the need for careful hyperparameter tuning. To address these challenges, various refinements were implemented, such as optimizing the code for efficiency, carefully reviewing and cleaning the data, and systematically exploring different model architectures and training strategies.

2.4 Implementation and Validation

The performance of the trained CNN model was rigorously assessed using the heldout test set, which comprised 20% of the total dataset. Two primary metrics were used for evaluation:

- **Sex Classification:** Accuracy was used to measure the model's ability to correctly classify sex from the ECGs. This metric represents the proportion of correctly classified instances (male or female) out of the total number of instances in the test set.

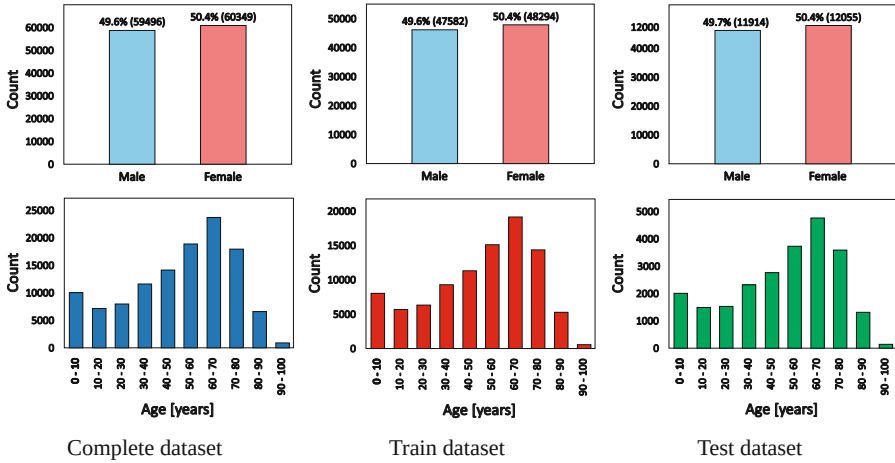


Fig. 3. Dataset demographics: Sex distribution (male/female) and Age distribution (decadal bins) for the complete dataset (column 1, blue for age), training (column 2, red for age), and testing (column 3, green for age) subsets.

- **Age Estimation:** Mean absolute error (MAE) was used to quantify the model's accuracy in predicting age. MAE calculates the average absolute difference between the predicted age and the actual chronological age for each individual in the test set. This metric provides a measure of the average prediction error in years.

To evaluate the model's practical applicability and potential for real-world use, it was integrated into the clinical systems of the *Instituto de Cardiología de Corrientes*. This integration involved:

- **Model Export:** The trained model was exported in a format that could be readily deployed and accessed within the clinical environment. This ensured that the model could be used for real-time predictions without requiring retraining.
- **API Development:** A RESTful API (Application Programming Interface) was developed to facilitate communication between the clinical system and the AI model. This API allowed the clinical system to send ECG data to the model and receive predictions in real-time. The API was designed to handle ECG data in PDF format, which was the standard format used in the clinical setting.
- **User Interface Integration:** The model's predictions were integrated into the existing user interface of the clinical system. This allowed healthcare professionals to view the predicted age and sex alongside other patient information, enabling them to assess the model's output in the context of the patient's overall clinical picture.
- **Real-time Testing:** The integrated system was tested in real-time using new ECGs from patients. This allowed for the evaluation of the model's performance on unseen data and its ability to provide accurate predictions in a clinical setting.

The implementation and validation process presented several challenges that required adaptations and refinements:

- **Data Variability:** ECGs from different electrocardiograph machines can exhibit variations in format, resolution, and signal quality. To address this, the signal extraction and preprocessing algorithms were adapted to handle ECGs from different brands and models of machines. This ensured that the model could be applied to a wider range of ECG data.
- **Anonymization:** Maintaining patient privacy was a crucial consideration. The anonymization procedures were carefully reviewed and updated to ensure that no identifying information was inadvertently included in the data sent to the model.
- **Integration with Existing Systems:** Integrating the AI model into the existing clinical workflow required close collaboration with IT specialists and healthcare professionals. This ensured that the model's output was presented in a user-friendly and clinically meaningful way.

The implementation and validation process provided valuable insights into the practical challenges of deploying AI models in clinical settings. The adaptations and refinements made during this process have enhanced the robustness and generalizability of the system, paving the way for its wider adoption in clinical practice.

3 Results

The CNN model demonstrated promising results in extracting age and sex information from the ECGs.

- **Sex Classification:** The model achieved an accuracy of 87% in classifying sex from the ECGs. This indicates that the model correctly predicted the sex of 87% of the individuals in the test set. A confusion matrix was used to visualize the model's performance, showing the distribution of true positive, true negative, false positive, and false negative predictions.
- **Age Estimation:** The model achieved a mean absolute error (MAE) of 9 years for age estimation. This means that, on average, the model's predictions for age were off by 9 years. To further analyze the age estimation results, a scatter plot was generated, comparing the predicted age with the actual age for each individual in the test set. This visualization revealed a general trend of agreement between the predicted and actual ages, with some deviations observed for individuals at the extremes of the age range.

To explore the potential clinical implications of the age estimation results, further analysis was conducted to investigate the relationship between the discrepancy between predicted age and chronological age and the prevalence of cardiovascular risk factors.

The analysis revealed a correlation between the difference between predicted age and chronological age and the presence of cardiovascular risk factors. Individuals whose predicted age was significantly higher than their chronological age tended to have a higher prevalence of these risk factors. This finding supports the hypothesis that the AI-derived age from ECGs may reflect biological aging and could potentially serve as a biomarker for cardiovascular risk stratification.

A confusion matrix was generated to illustrate the performance of the model in sex classification, showing the counts of true positives, true negatives, false positives, and false negatives (Fig. 4).

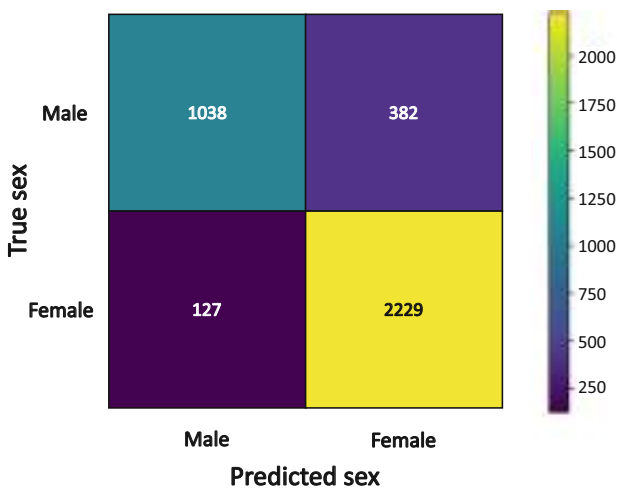


Fig. 4. Confusion matrix showing the performance of the model in sex classification.

A scatter plot was created to visualize the relationship between the predicted age and the actual age for each individual in the test set (Fig. 5).

Additionally, Fig. 6 presents the confusion matrix (Predicted Decade vs. Estimated Decade) to evaluate the ages estimated by decade. As can be seen, the errors are always located in cells adjacent to the main diagonal. This indicates that the model is more likely to make errors when estimating ages that are close to the true age. The largest errors are found in the first and last decades, where there are few data points. This suggests that the model may be less accurate when estimating ages in these decades.

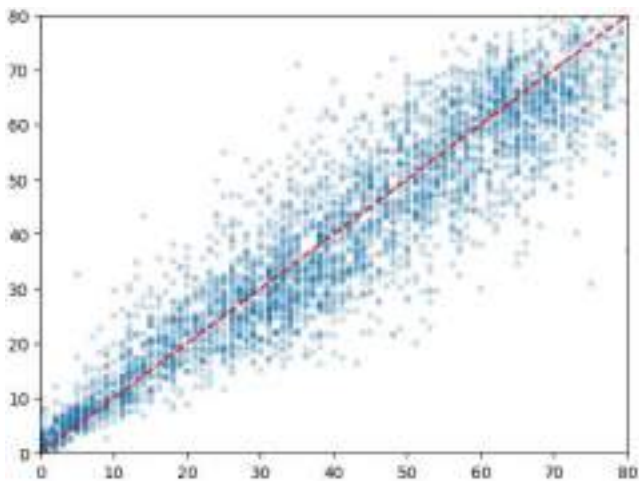


Fig. 5. Scatter plot showing chronological age (horizontal axis) vs. predicted age (vertical axis).

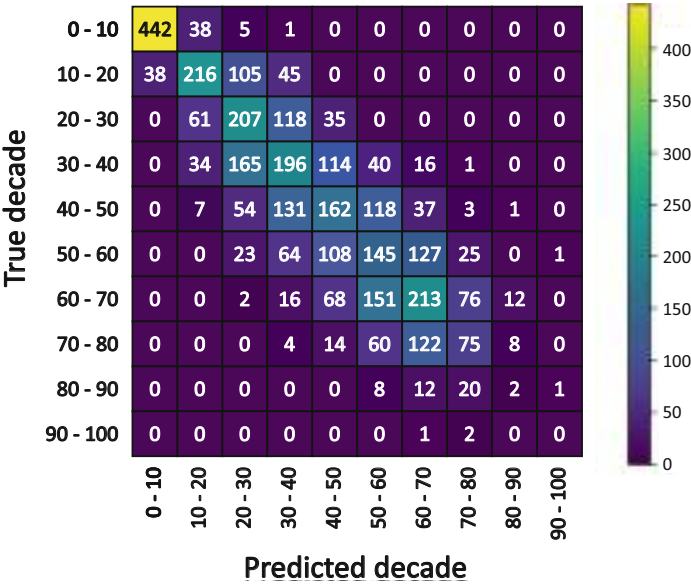


Fig. 6. Confusion matrix to evaluate the ages estimated by decade. True decade (vertical axis) vs. Predicted decade (horizontal axis).

These visualizations provide a clear and concise representation of the model’s performance and the key findings of the study.

4 Discussion

This study demonstrates the feasibility of estimating age and sex directly from ECG images using a CNN-based AI model. The model, trained on a large dataset of ECGs in PDF format, achieved promising results, with an accuracy of 87% for sex classification and a mean absolute error of 9 years for age estimation. These findings are consistent with previous studies that have explored AI-based age and sex estimation from ECGs, suggesting that our approach is comparable in performance to other state-of-the-art methods.

The novelty of our approach lies in its ability to extract relevant information from ECGs stored in digital image format (PDF), which is a common format for archiving medical records. This capability has the potential to unlock valuable insights from historical ECG data that may not be readily accessible in raw signal format. Furthermore, the model’s integration into the clinical workflow of the *Instituto de Cardiología de Corrientes* demonstrates its practical applicability and potential for realworld use.

The study also revealed an intriguing correlation between the discrepancy between predicted age and chronological age and the prevalence of cardiovascular risk factors. This finding suggests that the AI-derived age from ECGs may reflect biological aging and could potentially serve as a biomarker for cardiovascular risk stratification. Further

investigation is warranted to validate this hypothesis and explore the clinical utility of this novel biomarker.

The development and implementation of this AI model faced several challenges, including computational limitations, data variability, and the need for careful anonymization and integration with existing systems. These challenges were addressed through a combination of algorithmic adaptations, data preprocessing techniques, and close collaboration with IT specialists and healthcare professionals.

Some limitations of this study are:

- **Data source:** The study relied on ECGs from a single institution, which may limit the generalizability of the findings to other populations.
- **PDF format:** The use of ECGs in PDF format may introduce some limitations due to potential information loss during the conversion from raw signal data to images.
- **Computational constraints:** Computational limitations restricted the exploration of more complex AI models, such as LSTM and Transformer networks.

Future work will focus on:

- **Expanding the dataset:** Including ECGs from diverse sources to improve the generalizability of the model.
- **Exploring alternative AI architectures:** Investigating the use of LSTM, GRU, and Transformer networks to potentially enhance performance.
- **Refining feature extraction:** Developing more sophisticated signal processing and feature extraction techniques to capture subtle ECG characteristics.
- **Validating the age discrepancy biomarker:** Conducting further studies to validate the clinical utility of the AI-derived age as a biomarker for cardiovascular risk.

This study contributes to the growing body of research on AI applications in cardiology. The development of an AI model capable of estimating age and sex from ECG images has the potential to improve clinical decision-making, enhance risk stratification, and contribute to personalized medicine approaches in cardiovascular care.

Future work will focus on further improving the model's performance by exploring alternative architectures, such as LSTM, GRU, and Transformer networks, and by refining the data preprocessing and feature extraction techniques.

5 Conclusions

This study successfully developed and implemented an AI model capable of estimating age and sex from ECG images, demonstrating the potential of leveraging readily available clinical data for enhanced patient evaluation. The model's ability to extract information from ECGs in PDF format opens new possibilities for utilizing historical medical records and conducting retrospective analyses. The integration of the model into the clinical workflow of the *Instituto de Cardiología de Corrientes* highlights its practical relevance and potential for real-world impact. Furthermore, the observed correlation between AI-derived age and cardiovascular risk factors suggests a promising avenue for future research and the development of novel biomarkers for risk stratification. While challenges remain in terms of data variability, computational constraints, and the need

for further validation, this study represents a significant step towards the development of AI-powered tools for personalized cardiovascular care. The findings of this study contribute to the growing body of evidence supporting the use of AI in cardiology and give future innovations in this field.

Acknowledgments. We thank Fundación Sadosky for funding the project, and the Instituto de Cardiología de Corrientes for data access and collaboration. Special thanks to the IT and electrophysiology specialists who contributed to model implementation.

References

1. Jørgensen PG, Jensen JS, Marott JL, Jensen GB, Appleyard M, Mogelvang R (2014) Electrocardiographic changes improve risk prediction in asymptomatic persons age 65 years or above without cardiovascular disease. *J Am Coll Cardiol* 64(9):898–906. <https://doi.org/10.1016/J.JACC.2014.05.050>
2. Ladejobi AO et al (2021) The 12-lead electrocardiogram as a biomarker of biological age. *Eur Heart J Digital Health* 2(3):379. <https://doi.org/10.1093/EHJDH/ZTAB043>
3. Attia ZI, et al (2019) Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol* 12(9). <https://doi.org/10.1161/CIRCEP.119.007284/ASSET/740C6993-7105-406A-B1AB-CF0FE7743C20/ASSETS/IMAGES/LARGE/E007284FIG07.JPG>
4. Pignolo RJ (2023) AI-ECG and the prediction of accelerated aging. *Mayo Clin Proc* 98(4):502–503. <https://doi.org/10.1016/j.mayocp.2023.02.016>
5. Attia ZI, et al (2019) Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol* 12(9). <https://doi.org/10.1161/CIRCEP.119.007284>
6. Xu Y, Goodacre R (2018) On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test* 2(3):249–262. <https://doi.org/10.1007/S41664-018-0068-2/FIGURES/9>
7. Alowais SA et al (2023) Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 23(1):689. <https://doi.org/10.1186/S12909-023-04698-Z>
8. Alzubaidi L, et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8(1):1–74. <https://doi.org/10.1186/S40537-021-00444-8>
9. Beritelli F, Mienye ID, Swart TG, Obaido G, Jordan M, Ilono P (2025) Deep convolutional neural networks in medical image analysis: a review. *Information* 16(3):195. <https://doi.org/10.3390/INFO16030195>



Automatic Text Classification in Cardiac Risk Management: A Pilot Study

Jacopo Vitale¹(✉) , Mark C.H. de Groot² , Imo Höfer², Leandro Pecchia¹ ,
Saskia Haitjema² , and Bram Van Es²

¹ Research Unit of Intelligent Technology for Health and Wellbeing, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy
jacopo.vitale@unicampus.it

² Central Diagnostic Laboratory, Universitair Medisch Centrum Utrecht,
Heidelberglaan 100, 3584, CX Utrecht, Netherlands
<https://iht.unicampus.it/>

Abstract. This pilot study explores the feasibility of using text from Electronic Health Record (EHR) at the University Medical Center Utrecht to classify patient eligibility for cardiovascular risk management. The primary goal is to improve the identification of at-risk patients and facilitate timely, personalized interventions. Through a ResNet architecture, which is used as a feature extractor, the outputs are then passed into a neural network classifier to predict patient eligibility for cardiovascular risk management. Despite excluding numerical data from EHR during model training and inference, and the imbalanced nature of the dataset, the model achieved promising results with an accuracy of 85% and an F1-score of 0.85. This preliminary analysis demonstrates the feasibility of the approach and establishes a solid foundation for further improvements. Future research directions include integrating Explainable AI techniques to enhance model transparency, expanding the dataset, and addressing class imbalance through data augmentation or resampling techniques.

Keywords: EHR · NLP · Cardiovascular Risk

1 Introduction

1.1 Background

In the geriatric population, cardiovascular risk (CVR) represents a significant challenge due to the increased prevalence of cardiovascular diseases (CVDs) with advancing age [1,2]. Older adults often exhibit a complex clinical profile, where multiple factors including comorbidities, age-related physiological changes, frailty, and polypharmacy synergistically contribute to an elevated CVR [4]. Consequently, effective CVR management in this demographic necessitates an individualized, comprehensive approach that evaluates overall health status, functional capacity, and quality of life. Moreover, leveraging structured

clinical data from Electronic Health Records (EHR) offers a valuable opportunity for the early identification and continuous monitoring of high-risk individuals, thereby enabling the implementation of targeted preventive and therapeutic strategies [3]. The University Medical Center Utrecht (UMCU) possesses a rich repository of patient data within its EHR system. This data, encompassing a wide range of patient demographics, medical history, diagnoses, treatments, and laboratory results, presents a valuable opportunity to enhance CVR management. Patients at UMCU follow a well-defined clinical pathway, which may begin with a phone call or an in-person visit. Each patient's medical record is assigned an identification code based on the type of appointment or visit, without considering physiological parameters. Among the twenty-six possible codes that can be assigned, four designate the automatic enrolment to the CVR management programme. This eligibility assignment system has significant limitations, as it does not account for the patient's physiological parameters or clinical history. By leveraging the power of deep learning (DL) techniques, the UMCU aims to develop models capable of accurately classifying patients who would benefit most from targeted CVR management interventions. This initiative has the potential to improve patient outcomes by enabling earlier identification of at-risk individuals and facilitating timely, personalized care.

1.2 State of the Art

DL is transforming cardiovascular risk management by leveraging the wealth of data stored in EHRs, utilizing longitudinal health measurements, clinical notes, and other data types to enhance risk prediction and improve patient outcomes. By analyzing complex and irregular data patterns, DL models often outperform traditional risk assessment tools, incorporating a wider range of predictors and handling missing or inconsistent data more effectively [5].

A key application of DL in this domain is natural language processing (NLP), which enables the extraction of crucial clinical information from unstructured text within EHRs, such as physician notes. Transformer-based models like BERT have demonstrated high accuracy in identifying risk factors for heart disease from clinical documents, allowing for a more comprehensive understanding of patient health and disease progression [6]. Furthermore, DL facilitates the integration of structured and unstructured data in multimodal approaches, leading to more effective cardiovascular risk stratification and enabling personalized treatment strategies [7,8].

This pilot study explores the feasibility of using deep learning to assess patient eligibility for cardiovascular risk management (CVRM) programs based solely on outpatient clinic consult notes. These consult texts are extensive and contextually dense, containing a mix of relevant and extraneous information, which poses a challenge for accurately extracting key clinical elements. NLP techniques are applied to process these lengthy consults, identifying essential patient risk factors while filtering out non-informative content, ensuring that DL models can effectively leverage unstructured text to support risk stratification and personalized care.

Despite its potential, several challenges must be addressed for widespread adoption. The performance of DL models heavily depends on high-quality, complete data, as incomplete or poorly structured EHR data can hinder accuracy and reliability. Additionally, bias in training data remains a significant ethical concern, as it can perpetuate existing health disparities [9]. Addressing these limitations is crucial for the effective deployment of DL-based solutions in cardiovascular risk management.

2 Materials and Methods

This pilot exploration investigates the feasibility of using text from EHR consult letters at the UMCU to classify patient eligibility for cardiovascular risk management.

2.1 Dataset

In this pilot all patients who are eligible for the programme, i.e. having one (or more) subgroup codes, are labelled as 1 (eligible), otherwise 0 (not eligible). The dataset consists of 4431 consult letters. This initial exploration focuses on processing the textual data within these consult letters using a deep learning approach. Patients data has been divided using an 80 – 20 split ratio, as detailed in Fig. 1.

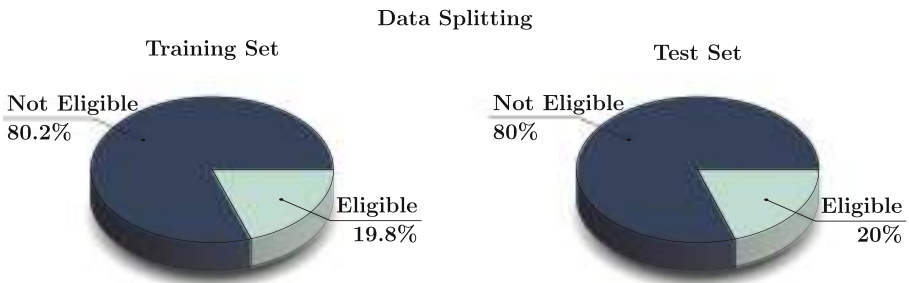


Fig. 1. Data is splitted keeping constant class samples ratio

2.2 Model

The methodology involved the following steps (Fig. 2): first, a custom Byte Level Tokenizer [10] was trained on all available consult letters written in Dutch. A Byte-Level Tokenizer processes text by converting each character into its raw byte representation, without relying on language-specific rules like spaces or punctuation. It treats all characters, including special symbols, as distinct tokens

in their byte form. This method ensures universal compatibility and simplifies handling of diverse text inputs, making it effective for multilingual and mixed-language scenarios. To ensure adequate representation of the majority of the data, a vocabulary size of at least 5000 tokens was found to be necessary. Pre-trained tokenizer was then used to convert the text of each consult letter into a sequence of numerical tokens.

These tokenized sequences were then passed through an embedding layer to create vector representations of each token, embedding dimension is set to 256. Embedded tokens are then concatenated together in a matrix 5000×256 given as input to the Residual Network (ResNet) [11]. Here, the ResNet it is used as a feature extractor to capture relevant patterns and relationships within the textual data. Finally, the extracted features were flattened and passed into a neural network classifier to predict patient eligibility for cardiovascular risk management (0 or 1).

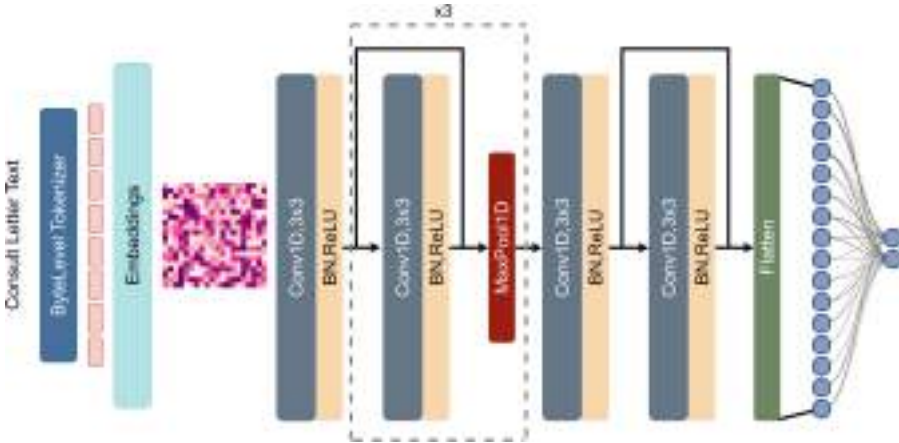


Fig. 2. Methodology workflow

3 Results and Discussion

Given the highly imbalanced nature of the dataset, accuracy alone may not fully reflect the model’s performance. Therefore, the F1-score and Matthews Correlation Coefficient Score (MCC) were considered more reliable indicators, as they account for both class imbalance and the trade-off between precision and recall. Results metrics are summarised in Table 1 and Fig. 3.

Given that the model is trained on manually annotated labels (as detailed in Sect. 2.1), particular emphasis is placed on analyzing the model’s false positive (FP) and false negative (FN) outputs. Since the automated classification method relies solely on consultation letters, it is imperative to investigate the

Table 1. Model Metrics summary on the test set

Metric	Value
Accuracy	85%
F1-Score	0.85
Precision	0.85
Matthews Correlation Coefficient	0.49

Actual	Not Eligible	655	56
	Eligible	80	95
	Predicted	Not Eligible	Eligible

Fig. 3. Confusion Matrix on the Test Set

discrepancies whereby the model classifies patients as eligible for the program despite their manual labels indicating otherwise and vice versa. These findings suggest that manual eligibility assessments possess inherent limitations, which may partly account for the high dropout rate observed at UMCU during patient enrollment. Furthermore, the automatic eligibility assessment not only has the potential to enhance patient quality of life, but it also represents a cost-effective health technology worthy of further evaluation. Notably, the deep learning-based improvements offer substantial economic benefits, as the cost of a cardiovascular risk management program is significantly lower than that of treating cardiovascular events.

4 Conclusion

This preliminary analysis, conducted in collaboration with UMCU, demonstrates the potential of using a ResNet-based model for binary text classification in EHR. In this pilot model, numerical data from the EHR were not utilized in the model's training or inference processes. Instead, these data were subjected to exploratory data analysis to better understand their distribution and potential relevance for future integration into the framework. Despite the exclusion of numerical data and the challenges posed by the imbalanced dataset, the model achieved promising results. These findings highlight the feasibility of the approach and establish a solid foundation for further improvements. Furthermore,

incorporating Explainable AI (XAI) techniques will be a priority to increase transparency and provide insights into the factors influencing the model's predictions. Expanding the dataset and addressing class imbalance through techniques such as data augmentation or resampling are also planned to improve the robustness and generalizability of the approach.

Funding Information. The work was supported in part by the ODIN Project, as part of the European Union's Horizon 2020 Research and Innovation Program, under Grant 101017331. **Declarations**

Ethics approval and consent to participate. The UMCU quality assurance research officer confirmed under project number 22U-0292 that this study does not fall under the scope of the Dutch Medical Research Involving Human Subjects Act (WMO) and therefore does not require approval from an accredited medical ethics committee. The study was performed compliant with local legislation and regulations. All patient data were deidentified in compliance with the European Union General Data Protection Regulation, and as a result, written informed consent was not required by the UMCU ethical committee.

Competing interests. The authors declare that they have no competing interests.

References

1. World Health Organization: Cardiovascular diseases (CVDs) Fact Sheet. [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
2. Tsao, C.W., et al.: American heart association council on epidemiology and prevention statistics committee and stroke statistics subcommittee: heart disease and stroke statistics-2023 update: a report from the american heart association. *Circulation* **147**(8), e93–e621 (2023). <https://doi.org/10.1161/cir.0000000000001123>
3. Hoyt, R.E., Yoshihashi, A.K.: Health informatics: practical guide for healthcare and information technology professionals. Inf. Educ. (2014)
4. Lakatta EG, Levy D (2003) Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: part i: aging arteries: a “set up” for vascular disease. *Circulation* **107**(1):139–146. <https://doi.org/10.1161/01.CIR.0000048892.83521.58>
5. Li, C., et al.: Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records. *Eur. Heart J. Digit. Health.* **5**(1), 30–40 (2023). <https://doi.org/10.1093/ehjdh/zta058>
6. Houssein EH, Mohamed RE, Ali AA (2023) Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci. Rep.* **13**(1):7173. <https://doi.org/10.1038/s41598-023-34294-6>
7. Bagheri, A., Groenhof, T.K.J., Veldhuis, W.B., de Jong, P.A., Asselbergs, F.W., Oberski, D.L.: Multimodal learning for cardiovascular risk prediction using EHR data. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (2020). <https://doi.org/10.48550/arXiv.2008.11979>

8. Singh, M., et al.: Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review. *EClinicalMedicine*. **73** (2024). <https://doi.org/10.1016/j.eclinm.2024.102660>
9. Overmars LM et al (2022) A wolf in sheep's clothing: reuse of routinely obtained laboratory data in research. *J. Med. Internet Res.* 24(11):e40516. <https://doi.org/10.2196/40516>
10. Kozma, L., Voderholzer, J.: Theoretical Analysis of Byte-Pair Encoding. *arXiv preprint* (2024). <https://doi.org/10.48550/arXiv.2411.08671>
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778. IEEE Press, New York (2016). <https://doi.org/10.1109/CVPR.2016.90>



Genetic Algorithm for Predicting Primary Imatinib Resistance in Locally Advanced or Metastatic Gastrointestinal Stromal Tumors

Martina Sassi^{1,2}(✉) , Jacopo Vitale¹ , Maria Elisabetta Pagnano¹ ,
Laura Risi Ambrogioni², Margherita A. G. Matarrese¹ , Bruno Vincenzi² ,
and Leandro Pecchia¹

¹ Research Unit of Intelligent Technology for Health and Wellbeing, Department of Engineering, Università Campus Bio-Medico di Roma, Via Alvaro del Portillo, 21, 00128 Rome, Italy

martina.sassi@unicampus.it

² Fondazione Policlinico Universitario Campus Bio-Medico di Roma, Via Alvaro del Portillo, 200, 00128 Rome, Italy

Abstract. Gastrointestinal stromal tumors (GISTs) are the most common mesenchymal neoplasms, primarily driven by KIT or PDGFRA mutations. While tyrosine kinase inhibitors (TKIs), such as imatinib, have transformed treatment, resistance remains a significant challenge. Alternative TKIs offer options but with increased toxicity and inconsistent efficacy. This study investigates artificial intelligence-driven radiomics for predicting primary TKI resistance in GIST patients. Sixteen patients were classified as imatinib-sensitive or -resistant based on CT scans taken at baseline and one year post-treatment. Expert radiologists manually segmented CT scans and extracted the volume of interest, composed of the principal mass and metastasis. A total of 93 radiomic features were extracted from each axial slice from the axial view. A genetic algorithm was implemented for feature selection and model hyperparameter optimization. The Support Vector Classifier demonstrated the highest accuracy (95%), underscoring the potential of AI-driven radiomics to guide personalized GIST treatment.

Keywords: artificial intelligence · genetic algorithm · gastrointestinal stromal tumors · machine learning · radiomics

1 Introduction

Gastrointestinal stromal tumors (GISTs) represent the most prevalent mesenchymal neoplasms, with an estimated annual incidence of approximately 10-15 cases per million individuals [1]. These tumors originate from the interstitial cells of Cajal or their precursors and are predominantly driven by activating mutations in the KIT or PDGFRA genes, which play a crucial role in tumor initiation, progression, and response to targeted therapies [2]. Approximately 85%

of GISTs harbor such mutations, with KIT mutations occurring in 75-80% of cases and PDGFRA mutations in 5-10% [3]. A subset of GISTs lacks mutations in both KIT and PDGFRA and is characterized by distinct oncogenic mechanisms, including succinate dehydrogenase (SDH) deficiency, or may harbor mutations in BRAF or NTRK, highlighting their unique molecular landscape and potential resistance to conventional tyrosine kinase inhibitors (TKIs) [4, 5]. Given this molecular heterogeneity, comprehensive genomic profiling is essential for accurate disease characterization, prognostic stratification, and the selection of appropriate therapeutic strategies [6].

The advent of TKIs has revolutionized the management of GISTs. Imatinib, a first-generation TKI targeting KIT mutations, has significantly improved survival outcomes in patients with advanced disease [7]. However, resistance to imatinib remains a major challenge, necessitating the development of second- or third-line TKIs and beyond, such as Sunitinib, Regorafenib, and Ripretinib, designed to overcome specific resistance mechanisms [8]. Nevertheless, none of these alternative therapies have demonstrated superiority over Imatinib in terms of overall survival, progression-free survival, or drug-related toxicity [9]. Imatinib has remained the first-line treatment of choice over the years, not only due to its efficacy but also its favorable tolerability profile [6]. For GISTs harboring primary resistance mutations, such as PDGFRA D842V-mutant GISTs, specific inhibitors like Avapritinib have been developed to overcome primary imatinib resistance; however, these agents are associated with a significantly higher toxicity profile [10].

Even in patients with Imatinib-sensitive mutations, primary resistance or early disease progression may occasionally occur, often attributable to secondary mutations within the kinase domains of KIT or PDGFRA, leading to reduced drug sensitivity and necessitating alternative therapeutic strategies [7].

Beyond systemic therapy, local treatment modalities have emerged as critical components in managing disease progression, particularly in cases of oligoprogression. Surgical resection of progressing lesions, stereotactic body radiotherapy (SBRT), and interventional radiology techniques - including radiofrequency ablation (RFA), microwave ablation (MWA), and transarterial chemoembolization (TACE) - offer promising strategies to achieve local disease control while maintaining the efficacy of ongoing systemic therapy [6]. These interventions may prolong treatment duration, delay the transition to subsequent lines of therapy, and ultimately improve patient outcomes.

Recent advancements in artificial intelligence (AI) and radiomics have introduced innovative methodologies for predicting therapeutic resistance and assessing treatment response in GISTs. Radiomics enables the extraction of high-dimensional quantitative imaging features from computed tomography (CT) scans, facilitating a more detailed assessment of tumor heterogeneity and response patterns [6]. Machine learning algorithms can integrate radiomic signatures with clinical and molecular data to identify patients at elevated risk of primary resistance or early disease progression. This AI-driven approach holds

significant potential for refining treatment selection, personalizing therapeutic strategies, and enhancing clinical decision-making.

This study aims to investigate the role of AI and radiomics in predicting primary resistance to TKIs in patients with GIST, with the overarching goal of optimizing therapeutic management and improving clinical outcomes.

2 Materials and Methods

2.1 Study Design and Patients

A total of 47 patients with locally advanced or metastatic GIST were analysed retrospectively. The inclusion criteria were the following: (1) histological diagnosis of locally advanced or metastatic GIST; (2) age older than 18 years; (3) patients treated with imatinib as first line treatment, and (4) among the scans performed, availability of CT scans in DICOM format both prior to and one year following treatment. Patient age, gender, clinical presentation, diagnosis, treatment, tumour site, secondary tumour association, histopathological analysis, risk classification, and prognosis were recorded. Patients underwent multiphasic CT imaging using a 128-slice scanner. The response to targeted therapy in GIST is evaluated based on the Response Evaluation Criteria In Solid Tumors (RECIST) criteria v1.1. Based on treatment outcomes, patients were classified into two response categories: class 0 (imatinib-sensitive), including those patients with partial response or stable disease, and class 1 (imatinib non-responsive), corresponding to patients with progressive disease.

2.2 Volume Annotation and Features Extraction

Expert radiologists manually segmented the CT scans, delineating the volume of interest (VOI) for each patient, which included both the primary tumor and metastatic lesions (Fig. 1). A pre processing phase was performed to enhance data quality. The pixel intensity was computed for each VOI in every CT scan. Intensity values were filtered by retaining only those within the 1st and 99th percentiles to remove outliers and noisy voxels that might bias the subsequent radiomics analysis. Subsequently, all slices were normalized and discretized to ensure comparability across scans acquired with different bit depth [11].

A total of 93 features were computed using the PyRadiomics library: (1) 18 first-order intensity-based statistics, such as mean, 25th, 50th, 75th percentiles, skewness, kurtosis, intensity kurtosis, and intensity variance; (2) 24 derived from the gray level co-occurrence matrices (GLCM); (3) 14 derived from the gray level dependence matrix (GLDM); (4) 16 derived from the gray level run length matrices (GLRLM); (5) 16 derived from the gray level size zone matrix (GLSZM); (6) 5 derived from the neighboring gray tone difference matrix (NGTDM). All features were computed for each masked CT slice from the axial view of the CT scan.

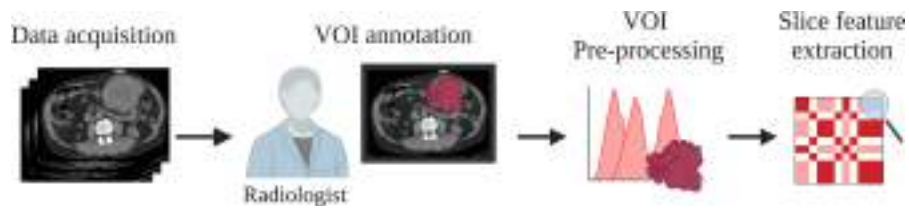


Fig. 1. Workflow of the image processing pipeline.

2.3 Classification Algorithm

Genetic algorithm (GA) is an evolutionary optimization technique that was inspired from the process of natural selection. Recently, GA has been used as a search strategy for dimensionality reduction, as well as for optimize the parameters of machine learning (ML) models [12]. In this research study, the GA has been used as a search strategy with different models: Support Vector Classifier (SVC), Random Forest Classifier (RC), Gradient Boosting (GB), and Extra Trees Classifier (ET) [13, 14].

The genetic algorithm goes through six important steps: initialization, selection, crossover, mutation, replacement, and termination. It operates on a population of N individuals, each represented by a genotype which consists of a set of binary genes. Each gene is encoded as an integer value (0 or 1), as shown in Fig. 2. Part of the genotype is responsible for features selection, while the remaining part defines model parameters. A specific gene combination corresponds to a unique phenotype, representing the function to be optimized (fitness).

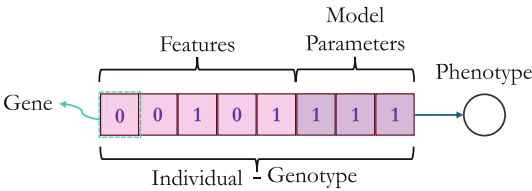


Fig. 2. Representation of an individual in the genetic algorithm. The genotype consists of binary genes, where one part is responsible for feature selection and the other for model parameter tuning. A specific gene combination defines the corresponding phenotype, which represents the function to be optimized (fitness).

The GA iteratively evolves the population by applying a sequence of genetic operators. The process begins with the initialization phase, where the GA initialize the population with individuals ($N=250$), each assigned a set of randomly generated genes. The selection phase entails evaluating the fitness of individuals. In this study, due to the unbalanced nature of the dataset, the Matthews

Correlation Coefficient (MCC) was employed as fitness. Subsequently, individuals are ranked and ordered based on their fitness value, from the highest to the lowest value (Fig. 3a). Crossover simulates the genetic recombination observed in nature, allowing pairs of selected individuals to exchange genetic information and generate a new offspring. In this study, parental pairing was performed between two randomly pooled individuals: parent I_1 from the top half of the sorted population and parent I_2 from the bottom half. In this research study, a single-point crossover was implemented, with a randomly chosen cut point (Fig. 3b). Offspring I_1^* is generated by inheriting the left hand side of the cut point from I_1 and the right segment from I_2 , while offspring I_2^* follows the opposite pattern. This process is repeated $\frac{N}{2}$ times, obtaining a new offspring of size N .

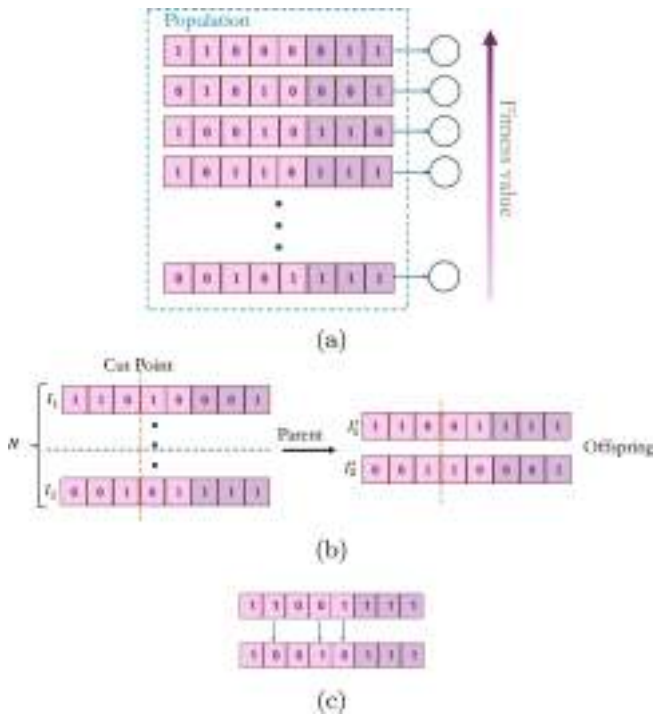


Fig. 3. Phases of the Genetic Algorithm: (a) Selection phase: individuals are evaluated on the fitness metric, ranked, and ordered from highest to lowest fitness value; (b) Crossover phase: parental pairing involves selecting I_1 from the top half and I_2 from the bottom half of the ranked population. A single-point crossover with a random cut generates offspring I_1^* and I_2^* by swapping segments; (c) Mutation phase: random changes by flipping selected gene values.

The mutation phase introduces random changes to the genetic information of each new individual. This process involves randomly selecting one or more

positions within the genetic information of an individual and flipping the values (0 to 1 or 1 to 0) at those positions (Fig. 3c). In this study, the mutation probability is dynamically adjusted over time following an exponential decay function. This approach ensures a high mutation probability at the beginning of the evolutionary process, promoting exploration of the solution space, while progressively reducing it in later stages to encourage convergence toward optimal solutions. The fitness of the new individuals is evaluated, and they are sorted to obtain the ranked offspring. During the replacement phase, a pairwise fitness-based comparison is conducted between the parent population and the offspring. For each individual I_i^* in the offspring, the corresponding individual I_i from the parent population is compared and eventually replaced to create the next population of best fitness values. This process is repeated for all N individuals, resulting in a new population of N individuals with the highest fitness values. The termination phase dictates the stopping criteria for the GA. The algorithm iterates through multiple generations, repeating the selection, crossover, mutation, and replacement phases until either the target fitness value is reached or a predefined number of epochs is exceeded ($G=100$). At the end of the process, the top-ranked individual is selected, whose genetic information represents the optimal combination of features and model parameters that maximizes the fitness function.

After optimizing and evaluating all feature-model combinations, SHapley Additive exPlanations (SHAP) analysis was performed to further interpret the model's predictions. SHAP values were computed to assess the contribution of each feature to the model's decisions, providing insights into the factors that most influenced classification outcomes. This approach enhances model explainability and helps identify the most relevant features in the decision-making process.

3 Results and Discussion

3.1 Cohort

Of the 47 total patients, only 16 patients (10 female, 6 male, age range 47–92, mean $\pm$ std age: 72.4 ± 13 years) with locally advanced or metastatic GIST fulfilled the inclusion criteria and were consequently included in the study.

The primary tumor site was gastric (56%), followed by ileal and duodenal locations, each accounting for 19%. Comprehensive molecular profiling, performed at the time of diagnostic biopsy or surgery, identified c-KIT exon 11 mutations in 10 patients (62%), a c-KIT exon 9 mutation in one patient, and no KIT/PDGFR mutations in two patients, while three had an unknown molecular status. All patients received imatinib at a daily dose of 400 mg without experiencing toxicity requiring treatment discontinuation. For treatment response assessment, CT scans acquired at baseline and one year after treatment initiation were analyzed. The median progression-free survival was 50 months (range: 5–129 months). After one year, 43.7% of patients had stable disease, 37.5% showed a partial response, and 18.8% experienced disease progression according

to RECIST v1.1 criteria. Based on treatment response, patients were classified in two classes: 13 patients in class 0, and 3 patients in class 1.

3.2 Classification Results

The final dataset comprises 1,284 slices, with 797 (60%) belonging to class 1 and 517 (40%) to class 0. A leave-three-subject-out (L3SO) validation strategy was employed to enhance the robustness of the evaluation. In each experiment, the test set included three different subjects, ensuring that all three minority-class subjects were represented across the different test sets. The remaining samples were randomly selected from the majority class for each experiment.

Table 1. Performance of the traditional classifiers.

Model	Metric	Test set 1	Test set 2	Test set 3	Mean
SVC	Accuracy(%)	93	99	92	95
	F1 score	0.9	0.99	0.9	0.93
	Sensitivity	0.9	0.99	0.83	0.91
	Precision	0.91	0.99	0.99	0.96
	MCC	0.85	0.98	0.84	0.89
RF	Accuracy(%)	77	100	90	87
	F1 score	0.59	0.99	0.88	0.78
	Sensitivity	0.45	1	0.8	0.7
	Precision	0.85	0.99	0.99	0.92
	MCC	0.49	0.99	0.81	0.72
ET	Accuracy(%)	84	100	88	91
	F1 score	0.76	1	0.86	0.87
	Sensitivity	0.7	1	0.75	0.82
	Precision	0.83	1	0.99	0.94
	MCC	0.65	1	0.78	0.81
GB	Accuracy(%)	79	95	91	88
	F1 score	0.63	0.94	0.89	0.82
	Sensitivity	0.5	0.92	0.8	0.74
	Precision	0.84	0.95	1	0.93
	MCC	0.52	0.9	0.82	0.75

A summary of the performances of the models is reported in Table 1. Among all models, the SVC demonstrated the highest performance, achieving an average accuracy of 95% and an F1-score of 0.93 across all test sets. Notably, the model exhibited its best performance on the test set 2, reaching 99% of accuracy, 0.98 of MCC, and 0.99 across of all other evaluation metrics. The other classifiers performed slightly worse but still achieved an average accuracy exceeding

87%, with F1-scores ranging from 0.78 to 0.87, recall between 0.70 and 0.82, precision between 0.92 and 0.94, and MCC ranging from 0.72 to 0.81. The RF classifier recorded the lowest mean accuracy (87%), along with lower F1-score and sensitivity compared to the other models, indicating a weaker generalization ability. The ET classifier outperformed the SVC on test set 2, achieving 100% of accuracy. However, overall, it obtained an average accuracy of 91%, an F1-score of 0.87, sensitivity of 0.82, precision of 0.94, and an MCC of 0.81.

Figure 4 presents the confusion matrices for the four models in their best-performing evaluations: SVC, RF, and GBs on test set 2, and ET on test set 3. The ET model achieved perfect classification, correctly identifying all test samples. Out of 214 test samples, the RF misclassified only one sample of class 0, while the SVC misclassified two instances - one from class 0 and one from class 1 - demonstrating its ability to correctly classify the majority of slices belonging to the samples. The GB model exhibited the highest number of misclassifications, despite achieving an average accuracy of 95%.

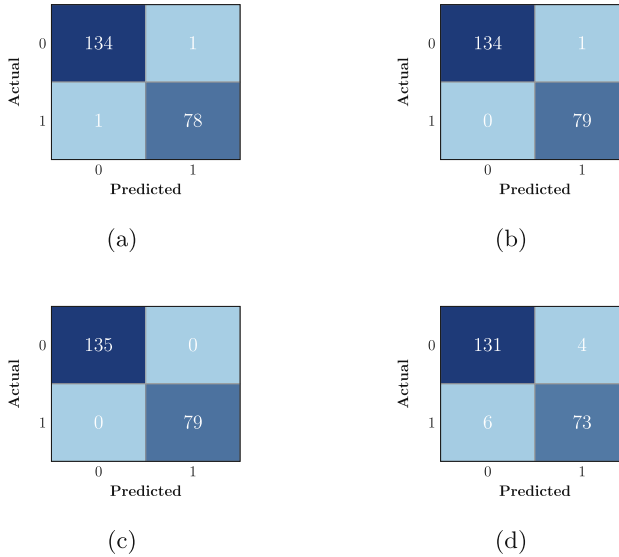


Fig. 4. Confusion matrix: (a) SVC; (b) RF; (c) ET; (d) GB.

To further interpret the model's decision-making process, SHAP analysis was employed to quantify the impact of each feature on the model's predictions. As an example, Fig. 5 illustrates the top 10 most important features ranked by how much they influenced the model's prediction for the SVC model on test set 2.

The feature importance is illustrated using a violin diagram, where the displayed features are ordered from highest to lowest based on their mean absolute SHAP values. The horizontal axis represents the SHAP value, where higher positive SHAP values correspond to a stronger positive contribution of the feature

to class 0 predictions, while negative values indicate a greater influence on class 1 classification. In this plot, each individual dot in each row represents the SHAP value for that particular feature for each sample of the dataset.

The color gradient represents the magnitude of the features values, shifting from red for higher values to blue for lower values. Inspection of the color distribution for each feature along the horizontal axis can provide global insights into the relationship between the feature values and its corresponding SHAP values. This visualization allows to discern how variations in the feature values directly influence the SHAP values, thereby revealing the model's sensitivity to different ranges of the feature.

In addition, the violin plot shows the distribution and density of SHAP values for each feature. A wider violin shape indicates greater variability in the feature's influence across different samples, suggesting that the impact of that feature may vary considerably among patients. Conversely, a narrower distribution implies a more consistent effect across the dataset.

For example, GLSZM Small Area Emphasis feature exhibited the highest SHAP value in Fig. 5, suggesting that this texture-based feature plays a crucial role in distinguishing between imatinib-sensitive and imatinib-resistant patients. This metric measures the proportion of small, homogeneous areas within the tumor region, and its negative SHAP values indicate a higher contribution to class 0 classification (imatinib- sensitive patients).

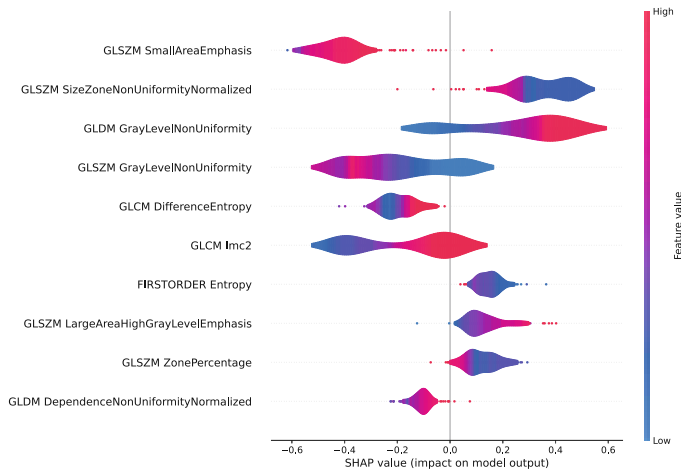


Fig. 5. SHAP violin plot for the SVC model on the test set. Colors represent feature values, while the distribution shows the impact of SHAP values on predictions.

The top 10 features identified across the three test sets were categorized into six distinct feature groups, and their percentage distribution within each category was analyzed. This evaluation offers valuable insights into the most predominant feature groups contributing to the classification task. Figure 6 presents

an example of this analysis, illustrating the distribution of the most important features identified by SHAP for the SVC model across the three test sets.

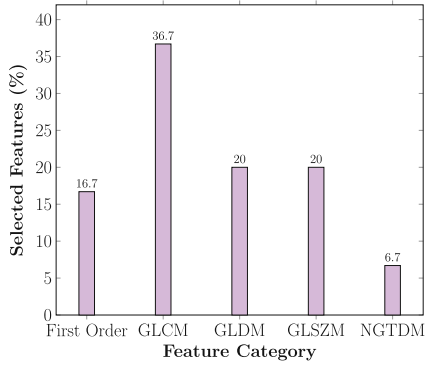


Fig. 6. Percentage distribution of the top 10 most important features identified by SHAP for the SVC model on the three test set, categorized based on their respective feature groups.

4 Conclusion

The integration of radiomics and artificial intelligence with clinicopathological data represents a promising approach for identifying patients at higher risk of imatinib resistance as early as the baseline CT scan. This technological synergy enables clinicians to enhance patient follow-up and detect relapses at an early stage, thereby significantly improving the effectiveness of therapeutic strategies. In this study, ML models were developed to differentiate GISTs with genotypes that are usually sensitive to imatinib (KIT mutations and PDGFR- α mutations different from exon 18 - D842V) from genotypes with lower sensitivity or that are insensitive to imatinib (PDGFR- α exon 18 - D842V mutation and WT GISTs) using features extracted from images of the axial view at the baseline CT scan.

Despite their potential relevance, these findings have certain limitations. First, the sample size is small and unbalanced due to the monocentric nature of the study. The inclusion of larger and multi-centre datasets might improve these promising preliminary results. Second, GISTs are often large tumors, and the manual segmentation of the entire volume is a time-consuming process that may introduce errors. The implementation of automated segmentation through deep learning algorithms could help mitigate this limitation in the future.

References

- Joensuu H, Hohenberger P, Corless CL (2013) Gastrointestinal stromal tumour. *The Lancet*. 382(9896):973–983. [https://doi.org/10.1016/S0140-6736\(13\)60106-3](https://doi.org/10.1016/S0140-6736(13)60106-3)
- Kawano K et al (1998) Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science*. 279(5350):577–580. <https://doi.org/10.1126/science.279.5350.577>
- Corless CL, Barnett CM, Heinrich MC (2011) Gastrointestinal stromal tumours: origin and molecular oncology. *Nat. Rev. Cancer* 11(12):865–878. <https://doi.org/10.1038/nrc3143>
- Rutkowski P, Seliga K, Debiec-Rychter M (2022) SDH-deficient gastrointestinal stromal tumors: disease models and therapeutic approaches. *Nowotwory. J. Oncol.* 72(5):326–333. <https://doi.org/10.5603/NJO.2022.00527>
- Miettinen M, Lasota J (2003) Gastrointestinal stromal tumors (GISTs): definition, occurrence, pathology, differential diagnosis, and molecular genetics. *Polish J. Pathol.* 54(1):3–24 PMID: 12817876
- Blay JY, Kang YK, Nishida T, von Mehren M (2021) Gastrointestinal stromal tumours. *Nat. Rev. Dis. Primers* 7(1):22. <https://doi.org/10.1038/s41572-021-00254-5>
- Heinrich MC et al (2008) Correlation of kinase genotype and clinical outcome in the North American intergroup phase iii trial of imatinib mesylate for treatment of advanced gastrointestinal stromal tumor: Calgb 150105 study by cancer and leukemia group b and southwest oncology group. *J. Clin. Oncol.* 26(33):5360–5367. <https://doi.org/10.1200/JCO.2008.17.4284>
- Antonescu CR et al (2005) Acquired resistance to imatinib in gastrointestinal stromal tumor occurs through secondary gene mutation. *Clin. Cancer Res.* 11(11):4182–4190. <https://doi.org/10.1158/1078-0432.CCR-04-2245>
- Napolitano A, Vincenzi B, Santini D, Tonini G (2015) Gastrointestinal stromal tumors: from a molecular understanding to clinical practice. *Fut. Oncol.* 11(16):2531–2544
- Li J, Zhang, X., et al.: Efficacy and safety of Avapritinib in treating Unresectable or metastatic gastrointestinal stromal tumors: a phase I/II, open-label, multicenter study. *The Oncology* **28**(2), 187–e114 (2023). <https://doi.org/10.1093/oncolo/oyac242>
- Yip SS, Aerts HJ (2016) Applications and limitations of radiomics. *Phys. Med. Biol.* 61(13):R150. <https://doi.org/10.1088/0031-9155/61/13/R150>
- Ghatasheh N, Altaharwa I, Aldebei K (2022) Modified genetic algorithm for feature selection and hyper parameter optimization: case of XGBoost in spam prediction. *IEEE Access*. 10:84365–84383. <https://doi.org/10.1109/ACCESS.2022.3196905>
- Breiman L (2001) Random forests. *Mach. Learn.* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Salim, Y., Rakasyah, A.F., Darwis, H., Manga, A.R.: Comparative analysis of machine learning algorithms and ensemble techniques for diverse image classification tasks. In: 19th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp. 1-7. IEEE (2025). <https://doi.org/10.1109/IMCOM64595.2025.10857517>



Echocardiographic Image-Based Classification of Cardiac Amyloidosis: A Proof of Concept

Maria Elisabetta Pagnano¹ (✉) , Martina Sassi¹ , Jacopo Vitale¹ ,
Margherita A. G. Matarrese¹ , Chiara Cirillo² , Filippo Crispino¹,
Giuseppe Limongelli² , and Leandro Pecchia¹

¹ Department of Engineering, Research Unit of Intelligent Technology for Health and Wellbeing, Università Campus Bio-Medico di Roma, Rome, Italy
m.matarrese@unicampus.it

² Inherited and Rare Cardiovascular Diseases Unit, Department of Translational Medical Sciences, University of Campania “Luigi Vanvitelli”, Naples, Italy

Abstract. Cardiac amyloidosis is a major cause of left ventricular hypertrophy, characterized by thickening of the left ventricular walls due to amyloid fibril deposition. Early diagnosis is crucial for optimizing patient management and reducing morbidity. This study proposes an approach based on automated echocardiographic image analysis for cardiac amyloidosis classification. A total of 21 subjects were included in this study. The A4-chamber view was selected from all cardiac chambers, and an automated pipeline was developed for field-of-view recognition. Machine and deep learning models were implemented to differentiate cardiac amyloidosis from healthy conditions using echocardiographic images. Preliminary results show high classification accuracy of 92% using Gradient Boosting, suggesting a potential improvement in differential diagnosis. This study highlights the growing role of artificial intelligence in cardiac imaging and paves the way for more efficient and precise diagnostic tools.

Keywords: Cardiac amyloidosis · artificial intelligence · classification · echocardiographic images

1 Introduction

Left ventricular hypertrophy (LVH) is a common clinical finding associated with adverse cardiovascular outcomes. LVH is characterized by the thickening of the left ventricular walls, and its main aetiologies include hypertension, hypertrophic cardiomyopathy, aortic valve stenosis, uraemia, cardiac amyloidosis (CA), and rare genetic conditions. Early diagnosis of LVH and accurate identification of its underlying cause are crucial for optimizing patient management and reducing morbidity, particularly in individuals under 60 years of age [1]. CA is a cardiomyopathy caused by the deposition of misfolded amyloid fibrils in the extracellular myocardial space. The most common subtypes are light-chain CA (AL-CA) and transthyretin (ATTR-CA). AL-CA represents a medical emergency, requiring rapid diagnosis and initiation of chemotherapy, whereas ATTR-CA can rapidly progress into symptomatic heart failure and is now amenable to medical treatment with transthyretin stabilizers [2].

This study serves as a proof of concept to explore whether AI image classification can support the identification of LVH and, more specifically, distinguish between CA (AL and ATTR) in echocardiographic images. To enhance the robustness of our analysis, we employed both machine learning (ML) and deep learning (DL) approaches. Traditional ML models can provide interpretability and leverage handcrafted features, while DL, particularly CNNs, can autonomously extract complex patterns from echocardiographic images. This complementary strategy allows us to explore different levels of feature abstraction and evaluate their effectiveness in distinguishing CA subtypes, ultimately aiming to improve diagnostic accuracy in clinical settings.

2 Material and Methods

2.1 Study Population

This study analyzed a dataset of echocardiographic images from 21 subjects who underwent standard transthoracic echocardiography at the Inherited and Rare Cardiovascular Disease Clinic of the University of Campania “Luigi Vanvitelli”-Monaldi Hospital (Naples, Italy) between March 2018 and December 2023, as part of routine clinical assessment.

All examinations were performed using a Vivid E9 ultrasound system (GE Healthcare, Horten, Norway) equipped with an M5S 3.5-MHz transducer, following standard chamber quantification protocols in accordance with current guidelines [3].

The study was approved by the Institutional Review Board of our university hospital, and all datasets were de-identified to ensure patient confidentiality. Patients were categorized into three groups: healthy subjects, those with a known clinical diagnosis of ATTR amyloidosis, and those with AL amyloidosis. The diagnosis of CA was established based on a combination of clinical evaluation, biomarker analysis, bone scintigraphy, and anatomic-histological findings.

2.2 Data Preprocessing

To ensure acquisition uniformity and enhance data comparability, we exclusively selected DICOM files containing the apical four-chamber (A4C) view, using a custom-built tool. This tool facilitates the selection process by displaying a preview of the first four frames of each DICOM file, allowing for a rapid and accurate identification of the A4C view. Despite all examinations being acquired in the same projection, significant variability was observed in the angular width of the field of view (FOV) across different DICOM files. This variation arose due to differences in echocardiographic acquisition settings. To address these inconsistencies and ensure a standardized region of interest, we implemented a two-phase preprocessing procedure.

Automatic FOV Recognition and Cone-Truncation Masking. The first preprocessing phase involved multiple steps to standardize the DICOM images. To obscure diagnostic annotations present at the image edges, black pixels were applied to the lateral margins of the ultrasound cone, specifically at the upper left, lower left, and along the right vertical border.

Additionally, the image reference system, initially positioned at the top left corner, was shifted to the vertex of the circular cone V (see Fig. 1a). The image was then systematically scanned from right to left to identify the first column containing non-black pixels. In cases where multiple rows exhibited non-null values, the lowest row within the lower half of the image was selected to ensure robust localization of the cone's geometric boundaries. Through this process, the lower right vertex (A) was identified, with its distance from vertex V corresponding to the radius (r) of the circular sector. The point M was identified as the projection of segment VA onto the y-axis (see Fig. 1b). The angular parameter was then computed through geometric analysis, with the angle α determined using the following formula [4]:

$$\alpha = \sin^{-1} \left(\frac{\overline{AM}}{r} \right) \quad (1)$$

The equation of the applied mask is:

$$M(\vartheta) = \begin{cases} 255, & \vartheta \in [\pi/2 - \alpha; \pi/2 + \alpha] \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

This process allowed the application of a truncated cone masking technique, effectively isolating the region of interest within each DICOM image (see Fig. 1c), regardless of variations in acquisition angle across subjects.

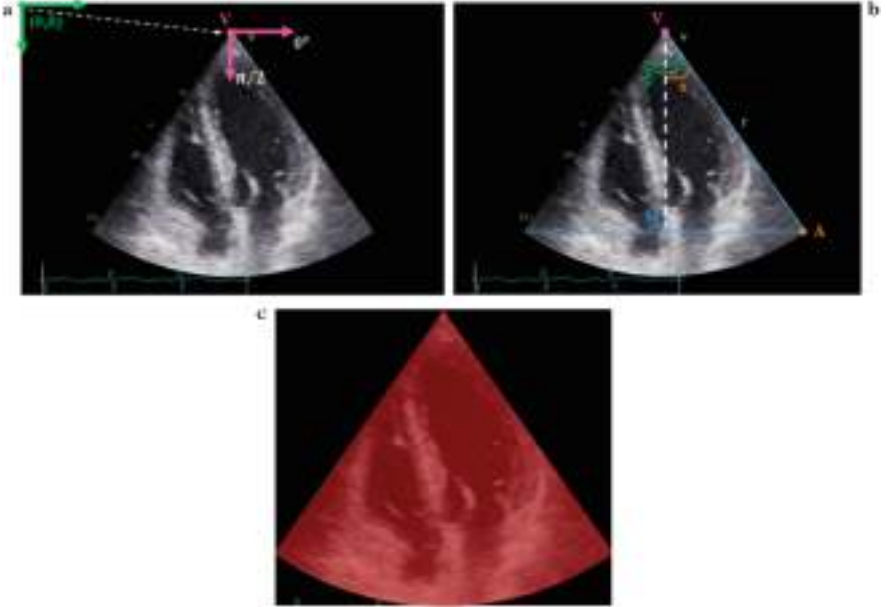


Fig. 1. Preprocessing steps: (a) Reference shift in V, (b) Geometric analysis with key points (A, M), radius, and angles, (c) Circular sector extraction.

Uniform Resizing of Images. Following the cone-truncation masking, the DICOM images exhibited variable dimensions due to differences in angular coverage. To establish a standardized reference size and ensure consistency across the dataset, a second preprocessing phase was implemented, involving the calculation of the maximum frame dimensions:

$$\max_{i=1,\dots,N} 2 \cdot \overline{AM}_i \tag{3}$$

where N is the total number of DICOM files. This computed size served as a reference for the padding step, enabling the definition of a standardized frame applicable to all images. This process ensured dataset uniformity, making it suitable for subsequent analysis. Maintaining consistent image dimensions is crucial for the effective application of both ML and DL techniques in later stages of the study.

2.3 Dataset Splitting

The dataset comprises 21 subjects and a total of 1515 frames. To ensure a robust model evaluation, data were split at the subject level, assigning all frames from each individual exclusively to either the training or test set. This strategy prevents data leakage and ensures that models are assessed on entirely unseen subjects, promoting better generalization and reducing the risk of overfitting to specific subject characteristics. Specifically, five subjects were allocated to the test set: two healthy and three affected by amyloidosis (AL and ATTR), while the remaining subjects were used for training. This partitioning ensured that 80% of the data was used for models training and 20% for testing their performance on unseen data.

In the ML approach, where feature extraction was performed on a subset of 10 frames per subject, the training set comprised 50 normal frames and 110 amyloidosis frames, while the test set included 20 normal frames and 30 pathological frames, for a total of 50 test frames.

In contrast, for the DL approach, all frames from each subject’s DICOM file were utilized, resulting in a test set of 477 frames and a training set of 1515 frames.

A summary of the dataset composition is provided in Table 1.

Table 1. Summary of dataset composition and splitting, highlighting the number of frames assigned to each set.

Approach	Set	Normal frames	Pathological frames	Total frames
Machine Learning	Training	50	110	160
	Test	20	30	50
Deep Learning	Training	329	1186	1515
	Test	150	327	477

2.4 Classification Models

Traditional ML techniques, which rely on handcrafted feature extraction, offer interpretable models with lower computational demands. Conversely, DL methods leverage neural networks to automatically learn hierarchical features from image data, often achieving superior performance at the cost of increased computational complexity and data requirements. In this study several ML and DL techniques were explored.

To provide a structured comparison, Fig. 2 illustrates the overall pipeline of both ML and DL methodologies, highlighting the key steps in feature extraction, model training, and classification.

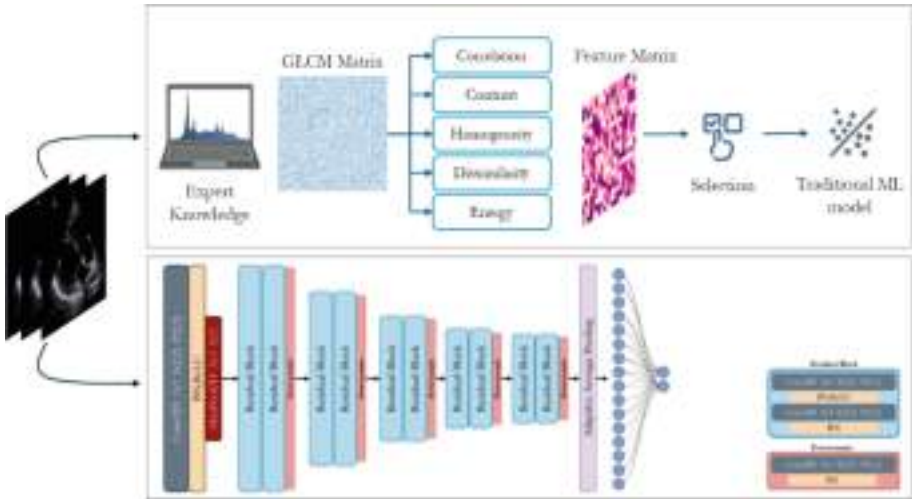


Fig. 2. Overall pipeline of machine learning and deep learning methodologies.

Machine Learning. For the ML approach, feature extraction was performed using texture analysis on the pre-processed and masked frames. Specifically, we employed the Gray-Level Co-occurrence Matrix (GLCM), a well-established method for quantifying texture features in grayscale images [5]. Each frame was first converted to grayscale, ensuring that pixel values were represented as 8-bit integers. The GLCM was then computed using predefined pixel distances (1, 2, and 3 pixels) and orientations (0, $\pi/4$, $\pi/2$, and $3\pi/4$ radians), capturing spatial relationships between pixel intensities across different scales and directions. From the GLCM, five key statistical features commonly used in texture analysis were extracted: Contrast, Dissimilarity, Homogeneity, Energy, Correlation. These features were computed for each combination of distance and direction, resulting in a total of 60 features (5 statistical measures $\times$ 3 distances $\times$ 4 orientations). Since features extracted from multiple frames of the same DICOM tend to exhibit high similarity, a subset of the most representative frames was selected. To ensure robust feature representation while minimizing redundancy, we extracted features from 10 key

frames per DICOM per subject, providing a more comprehensive yet efficient characterization of each case. The resulting feature vectors were then flattened and concatenated to represent the texture characteristics of each frame.

Due to the high correlation among the extracted features, Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving the most relevant information. The number of principal components was chosen to retain 98% of the total variance, ensuring a compact yet informative representation of the data.

The extracted feature vectors were used as inputs for several ML classifiers, including k-Nearest Neighbors (KNN), Random Forest (RF) and Gradient Boosting (XGBoost) [6, 7]. To optimize model performance, a grid search algorithm was used, which systematically explores all possible hyperparameter combinations through cross-validation. This exhaustive search evaluates each configuration using stratified K-Fold cross-validation ($k = 5$), ensuring a robust selection of optimal hyperparameters.

Deep Learning. In this case, all pre-processed and masked frames from the subject's DICOM file are considered. This ensures that the model leverages the full spatial information contained in the imaging data, potentially capturing more complex patterns and variations across frames. We employed ResNet18, a convolutional neural network (CNN) known for its efficiency and strong feature extraction capabilities in medical imaging [8]. ResNet18 was chosen due to its relatively compact architecture, which balances performance and computational efficiency, making it suitable for datasets of limited size. The residual connections allow for deeper network training without suffering from vanishing gradients, improving feature learning stability. Before being used as input to the model, all frames were resized to 224×224 pixels to match the ResNet18 input size.

2.5 Performance Evaluation Metrics

After training and optimizing our models, we evaluated their performance on the test set, which was held out during the initial dataset split. To ensure a comprehensive assessment, we computed the following classification metrics: Accuracy, Precision, Recall, F1-score, and Matthews Correlation Coefficient (MCC).

Precision measures the proportion of correctly identified positive predictions among all predicted positives, reflecting the reliability of the model in classifying positive instances. Recall, on the other hand, quantifies the proportion of correctly detected positive instances out of all actual positives, indicating the model's sensitivity.

The F1-score provides a balanced measure between precision and recall, being their harmonic mean. It is particularly useful in cases of class imbalance, as it accounts for both false positives and false negatives.

Finally, the Matthews Correlation Coefficient (MCC) offers a more robust evaluation by considering all four elements of the confusion matrix (true positives, true negatives, false positives, and false negatives). Unlike Accuracy, MCC remains informative even when class distributions are unbalanced, making it a reliable indicator of overall classification performance.

These metrics were computed for both ML and DL, allowing for a direct comparison of their classification performance. By evaluating multiple statistical measures,

we ensured a robust and interpretable assessment of the models' ability to distinguish between normal and pathological frames.

3 Results and Discussion

The bar plot in Fig. 3 illustrates the performance comparison among the evaluated models based on five key metrics.

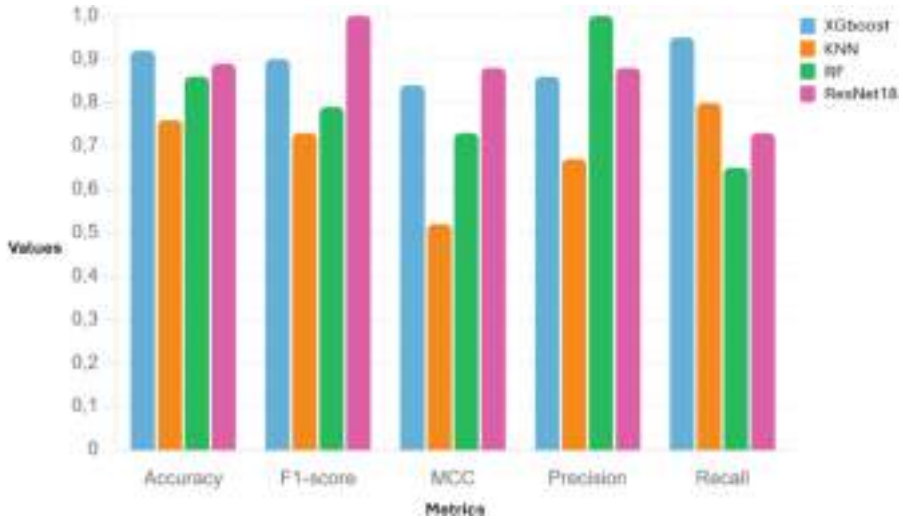


Fig. 3. Accuracy, F1-Score, Matthews Correlation Coefficient, Precision and recall of the models

Gradient Boosting achieved the highest overall performance, with an accuracy of 92%, F1-score of 0.90, and MCC of 0.84, indicating its superior ability to distinguish between normal and pathological frames. Additionally, its recall of 0.95 suggests a strong capability in identifying pathological cases, although its precision (0.86) is slightly lower than that of other models. Random Forest achieves an accuracy of 86% and an F1-score of 0.79, with perfect precision (1.00) but a lower recall (0.65). ResNet18, the deep learning approach, demonstrated robust classification capability, achieving an accuracy of 89%, an F1-score of 1.00, and an MCC of 0.74. Notably, its balanced precision (0.89) and recall (0.89) indicate a strong trade-off between sensitivity and specificity, highlighting its ability to generalize well despite the dataset size.

KNN, in contrast, exhibited the lowest overall performance, with an accuracy of 76%, an F1-score of 0.73, and the lowest MCC (0.52). Its precision of 0.67 and recall of 0.80 indicate a tendency to misclassify normal cases, reinforcing its limitations in handling complex echocardiographic features. These findings highlight the effectiveness of ensemble-based models, particularly Gradient Boosting and Random Forest, in leveraging handcrafted texture features for classification. Meanwhile, ResNet18 demonstrated competitive performance, suggesting that deep learning approaches can offer a viable alternative when properly fine-tuned.

To further analyze the classification behavior of the models, Fig. 4 presents the confusion matrices of Gradient Boosting (best ML model) and ResNet18. These matrices illustrate the number of correctly and incorrectly classified cases for each class, providing additional insight into their predictive reliability.

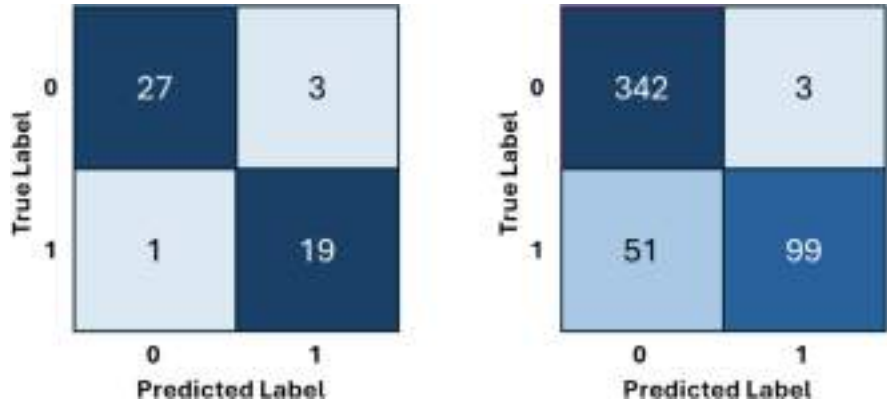


Fig. 4. Confusion matrices on the test set: XGBoost (left) and ResNet18 (right).

Specifically, the confusion matrix of ResNet18 reveals that, among the three pathological subjects in the test set, the model misclassified only three frames from a single subject, predicting them as normal. This error corresponds to approximately 4% of the total frames for that subject, indicating that the model is highly reliable in detecting pathological cases. Notably, the other two pathological subjects were classified with 100% accuracy, further demonstrating the model’s effectiveness in identifying disease-positive cases. However, the model exhibited a higher misclassification rate for healthy subjects, with one subject having 11% of its frames incorrectly classified as pathological, while another was misclassified in 78% of its frames. This indicates a tendency toward false positives rather than false negatives, meaning that the model is more likely to incorrectly classify a healthy subject as pathological rather than failing to detect an actual pathological case. From a clinical perspective, this behavior is preferable, as misclassifying a pathological patient as healthy could have severe consequences, potentially leading to missed diagnoses and delayed treatments. In contrast, false positives may result in additional but non-critical follow-up examinations, leading to a small but non-significant economic burden. However, they do not pose any risk of missing the disease in its early phase, making this trade-off more acceptable in a medical setting.

4 Conclusion

The application of AI in cardiac imaging through echocardiography has demonstrated promising potential in enhancing diagnostic accuracy, efficiency, and overall patient care [9]. In particular, AI-assisted early identification of CA could play a crucial role in reducing diagnostic delays and minimizing patient drop-off in the diagnostic-therapeutic

pathway. This proof-of-concept study highlights the feasibility of an automated classification tool that may contribute to streamlining clinical workflows and improving disease recognition. A key aspect of this study was the pre-processing pipeline, designed to optimize image quality and standardize inputs across different acquisitions. In particular, we implemented an automatic FOV recognition system, allowing for customized masking of each echocardiographic frame based on the aperture used during image acquisition. Among the tested models, Gradient Boosting achieved the highest performance, with an accuracy of 92% and strong capability in distinguishing between normal and pathological frames. These results confirm the effectiveness of ensemble learning in leveraging handcrafted texture features for classification. However, the promising performance of ResNet18 suggests that deep learning approaches could offer a viable and scalable alternative, especially when trained on larger and more diverse datasets. A key limitation of the present study is the small sample size, which impacts the generalizability of the findings. Future developments will focus on expanding the dataset to include a larger and more diverse cohort, incorporating cases of amyloidosis with non-cardiac involvement, as well as other conditions that may be misdiagnosed as amyloidosis.

References

1. Cuspidi C, Sala C, Negri F, Mancia G, Morganti A, Italian Society of Hypertension (2012) Prevalence of left-ventricular hypertrophy in hypertension: an updated review of echocardiographic studies. *J Hum Hypertens* 26(6):343–349. <https://doi.org/10.1038/jhh.2011.104>
2. Garcia-Pavia P (2021) Diagnosis and treatment of cardiac amyloidosis: a position statement of the ESC working group on myocardial and pericardial diseases. *Eur Heart J* 42(16):1554–1568. <https://doi.org/10.1093/eurheartj/ehab072>
3. Lang RM et al (2015) Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American society of echocardiography and the European association of cardiovascular imaging. *J Am Soc Echocardiogr Official Publ Am Soc Echocardiogr* 28(1):1–39.e14. <https://doi.org/10.1016/j.echo.2014.10.003>
4. Herlin IL, Ayache N (1992) Feature extraction and analysis methods for sequences of ultrasound images. In: Sandini G (Ed.) *Computer vision — ECCV'92. ECCV 1992. Lecture notes in computer science* (Vol 588). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-55426-2_6
5. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern* 3(6):610–621. <https://doi.org/10.1109/TSMC.1973.4309314>
6. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
7. Salim Y, Rakasyah AF, Darwis H, Harlinda I, Manga AR (2025) Comparative analysis of machine learning algorithms and ensemble techniques for diverse image classification tasks. In: *Proceedings of the 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pp 1–7. Bangkok, Thailand. <https://doi.org/10.1109/IMCOM64595.2025.10857517>
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 770–778. Las Vegas, NV, USA. <https://doi.org/10.1109/CVPR.2016.90>
9. Cirillo C et al (2025) Artificial intelligence for left ventricular hypertrophy detection and differentiation on echocardiography, cardiac magnetic resonance and cardiac computed tomography: a systematic review. *Int J Cardiol* 422:132979. <https://doi.org/10.1016/j.ijcard.2025.132979>



Application of Artificial Intelligence in Psychiatric Neuroimaging: A Systematic Review

Ilona Karpiel¹(✉), Ewelina Sobotnicka¹, and Ana Starcevic²

¹ Łukasiewicz Research Network- Krakow Institute of Technology, ul. Zakopiańska 73,
30-418 Krakow, Poland

Ilona.karpiel@kit.lukasiewicz.gov.pl

² Laboratory for Multimodal Neuroimaging, Institute of Anatomy, Medical Faculty, University
of Belgrade, 11000 Belgrade, Serbia

Abstract. With the growing application of artificial intelligence (AI) in mental health research, there is an increasing need to evaluate its clinical utility in diagnosing, predicting, and treating psychiatric disorders. This systematic review examines the role of AI methods in individuals undergoing neuroimaging assessments (fMRI, MRI, EEG). A comprehensive literature search was conducted in the PubMed Medline database and Cochrane Library, following PRISMA 2020 guidelines. The final analysis included 20 studies published between January 2020 and January 2025.

The findings indicate that deep learning models, particularly convolutional neural networks (CNNs), demonstrate higher accuracy in classifying psychiatric conditions such as autism spectrum disorder (ASD) and schizophrenia compared to traditional machine learning approaches. Moreover, AI is increasingly being explored for predicting treatment response, particularly in depression and bipolar disorder, by integrating neuroimaging with behavioral and physiological data. Despite promising results, data heterogeneity, lack of standardized validation protocols, and ethical concerns remain significant challenges for clinical implementation.

This review highlights the strengths and limitations of AI applications in psychiatric neuroimaging and underscores the need for standardized methodologies, larger datasets, and ethical guidelines to ensure the reliability and clinical applicability of AI-driven diagnostic tools. These findings provide a foundation for future research aimed at optimizing AI integration into psychiatric practice.

Keywords: brain · AI · Machine learning · central nervous system disorders · clinical trials

1 Introduction

Psychiatric disorders pose a significant challenge in clinical practice due to their heterogeneous nature and the lack of objective biomarkers. Unlike many neurological conditions, psychiatric disorders do not typically cause clear neuroimaging alterations in individual patients. However, multimodal neuroimaging techniques—such as voxel-based

morphometry, magnetic resonance imaging (MRI), functional MRI (fMRI), positron emission tomography (PET), and electroencephalography (EEG)—have been widely used for both diagnostic and experimental purposes [1, 2]. Advances in genetics and neuroimaging have enabled researchers to identify potential imaging endophenotypes, which could serve as measurable biomarkers in psychiatric treatment protocols [3]. Despite these advancements, clinical diagnosis and treatment assessment in psychiatry still rely primarily on symptomatology rather than objective neurobiological measures [4]. Artificial intelligence (AI) has become an increasingly valuable tool in psychiatric research and clinical practice. AI-driven methods, such as machine learning and deep learning, have demonstrated their potential in analyzing complex neuroimaging data, detecting subtle patterns, and predicting treatment responses [5, 6]. While AI applications have transformed fields such as neurology and oncology, their integration into psychiatric neuroimaging remains relatively limited [7]. Studies have reported promising results, such as AI-based models predicting response to antidepressant treatment based on fMRI findings [8] and identifying biomarkers for schizophrenia using deep learning techniques applied to structural MRI [9]. However, these findings remain fragmented, and a comprehensive synthesis of AI applications in psychiatric neuroimaging is lacking.

This systematic review aims to address the following key questions: Are artificial intelligence methods effectively used to diagnose, predict, or treat mental illness in individuals undergoing fMRI, MRI, or EEG assessments? Although research on AI in psychiatry is expanding, many studies focus on isolated aspects of neuroimaging or AI applications rather than providing an integrated perspective on their combined clinical utility. There remains a lack of systematic reviews synthesizing these advancements in a way that can inform both research and clinical practice. This review aims to bridge this gap by analyzing the current state of AI-driven neuroimaging in psychiatry, highlighting its limitations, and identifying areas requiring further research. By synthesizing existing findings, this work may contribute to the development of more precise and individualized psychiatric care [10, 11].

2 Preliminary Exploration of AI Applications in Psychology

Before developing the final search strategy for the systematic review, an initial search was conducted to explore the available literature and gather preliminary insights into the use of artificial intelligence (AI) in psychology. This search aimed to refine the clinical question developed according to the PICOS framework. The initial search was exploratory and did not constitute a formal systematic review; rather, it served as a foundation for building a more targeted search strategy, helping define relevant inclusion and exclusion criteria for the final review.

To better understand how AI is applied in psychology, the specialized PsycARTICLES database [12] was selected for the preliminary search. The keyword “machine learning” was chosen, as machine learning is a widely studied subset of AI in the field of psychology. The initial search yielded 101 results (filter: Year: 2020–2025, Article Type: Review Article, Subjects: Physical Sciences). Of these, five articles were identified as potentially relevant to the application of machine learning in psychology. However, one

article was found to be unsuitable for further analysis because it focused on decision-making processes in pilots, which was unrelated to psychology, leaving four articles for further review.

This preliminary search highlighted a key issue: while the database returned 101 articles that included “machine learning” and “AI” in the title, the majority were not directly relevant to the specific clinical question being addressed. Despite containing terms like “AI,” “machine learning,” and “artificial intelligence” in their titles, many articles did not discuss the use of AI in diagnosing, predicting, or treating mental illness in individuals undergoing neuroimaging procedures such as fMRI, MRI, or EEG. As a result, most articles were discarded due to their lack of relevance to the focus of this review.

Out of the 101 articles retrieved, five were initially identified as potentially relevant. Upon closer examination, only four articles aligned with the study’s clinical question and were selected for further analysis. These articles provided valuable insights into the role of machine learning and AI in psychological research, particularly in relation to mental health diagnosis and treatment. This process of exclusion and refinement significantly influenced the development of the final search strategy for the systematic review.

It is important to note that this was an initial exploration, not a systematic review. The findings from this search helped shape the criteria for the subsequent stages of the review and were instrumental in refining the approach. Specifically, these preliminary results allowed the authors to focus on machine learning applications in psychology and highlighted the need for a more thorough search for relevant studies.

For example, one study [13] introduces important topics related to the population of individuals with mental disorders, interventions using AI and machine learning, and outcomes related to treatment effectiveness. However, the article focuses more on psychotherapy than on the application of AI in neuroimaging diagnostics (fMRI, MRI, EEG), which may require differentiation in the context of your question.

Article [14] doesn’t directly address neuroimaging or specific clinical outcomes like diagnostic accuracy. However, its focus on cost-sensitive machine learning provides a useful theoretical foundation for the development of AI interventions in psychological diagnostics and treatments, which could be explored further in systematic reviews or clinical trials.

Article [15] does not directly focus on individuals with or at risk of mental illness undergoing neuroimaging procedures such as fMRI, MRI, or EEG. However, its emphasis on psychological scale development could indirectly apply to this population when creating scales to assess mental health conditions that may also be studied through neuroimaging techniques. The article discusses the use of neural networks (a form of deep learning) for automatic item generation in psychological tests. While it does not specifically address the application of AI for diagnosing mental illness or predicting outcomes using neuroimaging methods, the use of neural networks can be extrapolated to improve the development of diagnostic tools, scales, or assessments. These tools could, in turn, be used alongside neuroimaging studies in psychiatry or psychology, enhancing diagnostic accuracy and treatment strategies. Although the article does not compare neural network-based item generation with traditional methods, one could infer that traditional

psychological testing and scale development approaches might be compared with AI-based methods. The potential advantages of AI techniques include increased efficiency, personalization, and scalability, all of which can be beneficial for mental health diagnostics and treatment predictions, particularly when integrated with neuroimaging data. While the primary focus of the article is on the development of psychological scales, the use of AI in this context could improve the diagnostic accuracy and predictive performance of psychological assessments. These advancements are crucial when studying individuals undergoing neuroimaging procedures. The automatic generation of more accurate or comprehensive psychological scales could ultimately lead to better treatment efficacy and more personalized care. The article does not present a clinical trial or systematic review but focuses on the methodological application of neural networks in psychological research. The authors explain and apply the Psychometric Item Generator (PIG), an open-source natural language processing algorithm that produces large-scale, human-like, customized text output with a few mouse clicks. The PIG is based on GPT-2, a powerful generative language model, and runs in an interactive virtual notebook environment that executes code on state-of-the-art virtual machines at no cost. The authors demonstrate that the PIG is well-suited to generate large pools of face-valid items for novel constructs and create parsimonious short scales of existing constructs (e.g., Big Five personality traits) that perform strongly when tested in real-world conditions and benchmarked against current gold standards for assessment. This presents an effective, novel machine learning solution to an old psychological challenge.

Study [16] does not present a clinical trial or systematic review. Instead, it combines experimental methods with machine learning to explore the effects of ASMR. While not a systematic review or clinical trial itself, the methodologies described could inform future studies or systematic reviews that investigate the role of machine learning in understanding psychological phenomena, including in populations studied with neuroimaging techniques for mental health diagnoses. The article does not focus on mental illness or neuroimaging, but its use of machine learning to analyze psychological responses (in this case, ASMR) offers a potential framework for applying AI techniques to mental health diagnostics and treatment, particularly in combination with neuroimaging methods. The study utilized machine learning methods to categorize Autonomous Sensory Meridian Response (ASMR) content based on its sound structures. Four categories emerged, which were then used in a second study involving an online experiment ($N = 230$) to investigate the relationships between these categories of sound structures and participants' psychological experiences of ASMR. These findings help us understand how ASMR content generates the arousal and affect that audiences seek.

Machine learning algorithms are classified into supervised and unsupervised learning. The majority of the studies reviewed employed supervised learning, which is well-suited for classifying data and identifying patterns related to psychological outcomes. These studies demonstrate the growing relevance of machine learning techniques in psychology, particularly in the context of diagnostics and treatment personalization. The information gathered from this initial search played a crucial role in developing the search strategy for the systematic review. It enabled the authors to better understand the current landscape of AI and machine learning in psychology and to refine their approach for the later stages of the review process.

3 Materials and Method

3.1 Search Strategy and Information Sources

We systematically reviewed literature related to the question: Are artificial intelligence methods used to diagnose, predict mental illness or treat people with mental illness who also qualify for fMRI, MRI or EEG tests? The PICOS scheme was used to formulate the clinical question. The individual components of the PICOS scheme are defined as follows:

Population- Individuals with or at risk of mental illness undergoing fMRI, MRI, or EEG.

Intervention - Application of AI methods.

Comparator - Traditional diagnostic methods, non-AI approaches, or no comparator.

Outcome - Diagnostic accuracy, predictive performance, treatment efficacy, or other clinical outcomes.

Study Design - Systematic Review, Clinical Trial.

To create a search strategy, Dissemination for Systematic Reviews and PRISMA recommendations, we followed the Cochrane Handbook [17]. A complete search of the literature published in Polish or in English until January 2025 was conducted in PubMed Medline database and Cochrane. As part of the systematic review, we did not search the PsycARTICLES database because it contains mostly articles that are not systematic reviews and to which access is often restricted due to the need to pay fees.

The search strategy defines keywords, Boolean operators (AND, OR), filters (Free full text, Systematic Review), constraints (only Polish and English language), time span covered by the search (2020/1/1–2025/1/21)/date of last search (2025/1/21) and number of records found separately for each query (search queries) used in the search strategy, Table 1. The number of the searched results was reduced by increasing the sensitivity and specificity of the search. The criteria responsible for increasing sensitivity and specificity were: search for multi-word phrases in quotation marks, search for systematic reviews to obtain publications with a high degree of credibility, search for articles with full texts to gain an insight into the article's relevance, search for publications only in English and Polish, as these are the languages understood by the authors compiling this review.

The search strategy used in this systematic review involved comprehensive queries designed to capture relevant literature on AI applications in the diagnosis, prediction, and treatment of mental health conditions. The detailed search included the following steps. We employed a series of keywords and Boolean operators to capture relevant studies. The search query included combinations of mental health disorders, such as 'depression', 'schizophrenia', 'autism spectrum disorder', 'bipolar disorder', 'attention deficit hyperactivity disorder', and 'psychotic disorder', along with their corresponding Medical Subject Headings (MeSH) terms. These terms were connected using the Boolean operator 'AND' to ensure the inclusion of all relevant topics. For example:

- ((Depression) AND ("Depression"[Mesh])) OR (("Depressive Symptoms" OR "Depressive Symptom" OR "Symptom, Depressive"))
- ((Schizophrenia) AND ("Schizophrenia"[Mesh])) OR ((Schizophrenias OR "Schizophrenic Disorders" OR "Disorder, Schizophrenic"))

- Similarly for other disorders such as Autism Spectrum Disorder, Bipolar Disorder, ADHD, and Psychotic Disorders.

We included keywords related to artificial intelligence (AI) and neuroimaging techniques. For AI, terms like “Machine Learning”, “Deep Learning”, “Artificial Intelligence”, and “Computational Intelligence” were used. For neuroimaging, terms such as ‘fMRI’, ‘Magnetic Resonance Imaging’, and ‘EEG’ were included in the search query, alongside their corresponding MeSH terms:

- (“Machine Learning”[Mesh]) OR (“Artificial Intelligence”[Mesh])
- (“Magnetic Resonance Imaging”[Mesh]) OR (“Electroencephalography”[Mesh])

The search query also included terms related to diagnostic and clinical outcomes, such as ‘Diagnosis’, ‘Prediction’, ‘Treatment’, and ‘Prognosis’, to ensure a focus on the impact of AI applications in these areas:

- (“Diagnosis”[Mesh]) OR (“Prognosis”[Mesh]) OR (“Treatment”[Mesh])

We applied filters to narrow down the search to only include articles that were free full-text and systematic reviews. The time range for the search was from January 1, 2020, to January 21, 2025.

After performing the search in PubMed, we obtained 141 results. We then replicated this search strategy in the Cochrane Library with the same keywords and MeSH terms. This resulted in 37 trials that potentially matched the topic, Table 2. The selection of relevant data from these results was carried out in subsequent stages of the review process.

3.2 Criteria for Study Selection

The search for reliable scientific studies is complicated by imprecise and evolving terminology in the field. To address this challenge, the inclusion and exclusion criteria were carefully defined to ensure the selection of studies most relevant to the clinical question posed in our review. These criteria are outlined below.

Inclusion criteria:

- Studies utilizing digital technologies to enhance the clinical care of depression.
- Studies involving EEG in the adult ADHD population.
- Research focused on neurofeedback (NF) based on electroencephalography (EEG).
- Studies involving both children and adults.
- Use of machine learning or statistical learning methods.
- Studies involving human subjects only.
- Studies addressing the treatment and/or diagnosis of psychiatric disorders, with a focus on identifying technologies applicable to clinical practice.
- Studies involving bipolar disorder.
- Use of deep learning models for the classification of psychiatric disorders.

Exclusion criteria:

- Studies involving patients undergoing rehabilitation after myocardial infarction.
- Studies on patients suffering from chronic pain or neuropathic pain.
- Studies that combine genetic data.

Table 1. Search Results in PubMed Databases. The table presents the number of results obtained from queries related to various mental disorders, imaging technologies, artificial intelligence, diagnosis, prediction, and treatment. The final number of results is provided after applying the relevant filters.

ID	Search Query	Results
1	Depression or “Depression”[Mesh]	621,001
2	(“Depressive Symptoms” OR “Depressive Symptom” OR “Symptom, Depressive”)	67,535
3	“Depressive disorders” or “Depressive Disorder”[Mesh]	122,307
4	Schizophrenia or “Schizophrenia”[Mesh]	163,315
5	(“Schizophrenias” OR “Schizophrenic Disorders” OR “Disorder, Schizophrenic”)	163,455
6	“Autism spectrum disorder” or “Autism Spectrum Disorder”[Mesh]	42,743
7	(“Autism Spectrum Disorders” OR “Autistic Spectrum Disorder” OR “Autistic Spectrum Disorders”)	14,999
8	“Bipolar disorder” or “Bipolar Disorder”[Mesh]	57,392
9	(“Bipolar Disorders” OR “Disorder, Bipolar” OR “Psychosis, Manic-Depressive”)	7,724
10	“Attention deficit hyperactivity disorder” or “Attention Deficit Disorder with Hyperactivity”[Mesh]	34,851
11	(“Attention Deficit Disorders with Hyperactivity” OR ADHD OR “Attention Deficit Hyperactivity Disorder”)	48,779
12	“Psychotic disorder” or “Psychotic Disorders”[Mesh]	58,698
13	(“Disorder, Psychotic” OR “Disorders, Psychotic” OR “Psychotic Disorder”)	6,545
14	“Biomedical image segmentation”	93
15	“Machine learning” or “Machine Learning”[Mesh]	103,062
16	“Learning, Deep”	366
17	“Artificial intelligence” or “Artificial Intelligence”[Mesh]	77,477
18	(“Intelligence, Artificial” OR “Computational Intelligence” OR “AI (Artificial Intelligence)”)	1,646
19	fMRI or “Magnetic Resonance Imaging”[Mesh]	685,950
20	(“Imaging, Magnetic Resonance” OR “Magnetic Resonance Image” OR “Image, Magnetic Resonance”)	5,062
21	EEG or “Electroencephalography”[Mesh]	202,564
22	(EEG OR Electroencephalogram OR Electroencephalograms)	206,541
23	Diagnosis or “Diagnosis”[Mesh]	11,533,660
24	(Diagnoses OR Diagnose OR “Diagnoses and Examinations”)	11,533,660

(continued)

Table 1. (continued)

ID	Search Query	Results
25	Prediction or “Prognosis”[Mesh]	1,891,903
26	(Prognoses OR “Prognostic Factors” OR “Prognostic Factor”)	2,258,300
27	Treatment or “Therapeutics”[Mesh]	13,246,791
28	(Therapies OR Treatment OR Treatments)	13,319,446
29	Combined results: (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11) AND (#12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22) AND (#23 OR #24 OR #25 OR #26 OR #27 OR #28) AND filters: Free full text, Systematic Review, 2020/1/1 - 2025/1/21	141

- Animal studies.
- Research using biological samples, such as fecal or blood samples.
- Studies involving virtual reality.
- Studies focused on brain tumors.
- Invasive studies.
- Studies on magnetic seizure therapy.
- Research on Parkinson’s disease or Alzheimer’s disease.
- Studies using transcranial direct current stimulation (tDCS).
- Studies focused on diabetes.
- Studies published in languages other than English or Polish.
- Research on Personal Activity Intelligence (PAI).

3.3 Data Selection

The review process, aimed at assessing whether the scientific reports found met the inclusion criteria for the analysis, was conducted in stages. The first stage involved selection based on titles and abstracts, followed by selection based on full-text publications. The selection of studies was carried out on the basis of the adopted inclusion and exclusion criteria, according to the defined PICOS scheme.

Studies in the systematic review were selected at all stages by two analysts working independently. Any disagreements were resolved by consensus.

The analysis provides information on the number of scientific reports available at each stage of the research search and selection process. The process of study selection and data extraction followed a systematic approach based on predefined inclusion and exclusion criteria, in accordance with the PICOS scheme. To ensure transparency and reproducibility of the selection process, we utilized a PRISMA flow diagram. Although PRISMA guidelines are commonly associated with scoping reviews, they are also widely used in systematic reviews to provide a clear and structured representation of the study selection process. The use of PRISMA in this review enhances methodological rigor by visually detailing the number of records screened, excluded, and included, along with the reasons for exclusion at each stage. The process leading to the final selection

Table 2. Search Results in the Cochrane Library. The table presents the number of records retrieved from the Cochrane Library for queries related to various mental disorders, imaging technologies, artificial intelligence, diagnosis, prediction, and treatment.

ID	Search Query	Results
1	Depression	31,297
2	Depressive disorders	357
3	Schizophrenia	3,800
4	Autism spectrum disorder	1,443
5	Bipolar disorder	1,444
6	Attention deficit hyperactivity disorder	1,663
7	Psychotic disorder	163
8	Biomedical image segmentation	1
9	Machine learning	1,844
10	Artificial intelligence	1,429
11	fMRI	1,936
12	Magnetic Resonance Imaging	10,461
13	EEG	2,999
14	Electroencephalography	1,787
15	Diagnosis	52,904
16	Prediction	5,585
17	Treatment	234,862
18	Treatment-resistant schizophrenia	19
19	Combined mental disorder queries	5,777
20	Combined imaging & AI queries	7,992
21	Combined diagnosis, prediction & treatment	257,974
22	Final results after applying all filters	37

is presented in the form of a diagram in line with PRISMA recommendations, Fig. 1. The diagram also includes the reasons for excluding studies at the selection stages. The main reasons for exclusion were measurement methods described in the papers other than those assumed in the inclusion criteria, medical conditions, genetic testing and drug therapies.

In the first stage of the analysis, the titles of 177 publications were reviewed. Based on the preliminary selection of titles, 84 publications were excluded. The main reasons for exclusion were the lack of relevance to the research topic and non-compliance with the inclusion criteria, such as the focus on AI in the context of diagnosing, predicting, or treating mental health disorders. Ninety-three publications remained and proceeded to the abstract review stage. After carefully reviewing the abstracts, 66 publications were excluded, primarily due to the lack of relevant data concerning the use of imaging

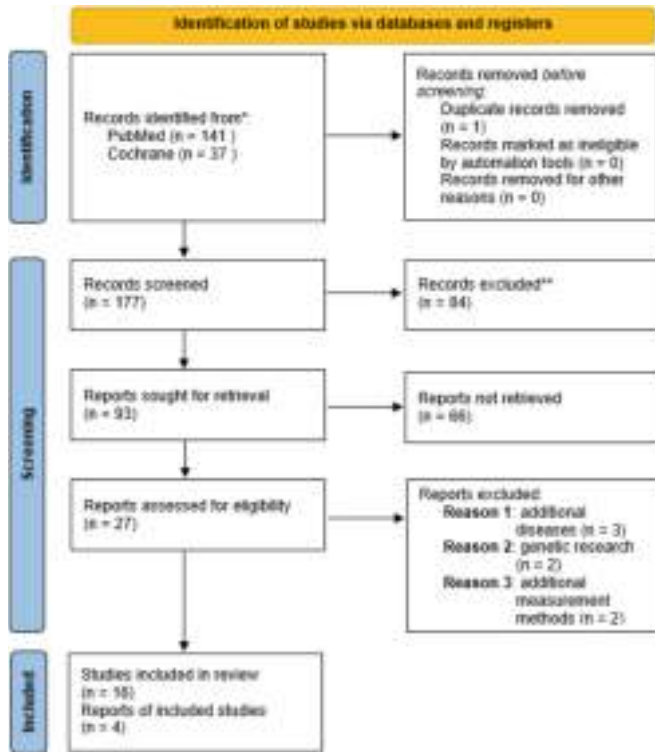


Fig. 1. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only. Flow chart illustrating the identification and screening process of studies via databases and registers. The process begins with 178 records identified from PubMed and Cochrane. After removing duplicates and ineligible records, 177 records are screened. Of these, 84 are excluded, leaving 93 reports sought for retrieval. 66 reports are not retrieved, and 27 reports are assessed for eligibility. 7 reports are excluded for reasons such as additional diseases, genetic research, and additional measurement methods. Finally, 16 studies are included in the review, with 4 reports of included studies. The chart is organized into sections: Identification, Screening, and Included.

technologies or AI, or an inappropriate clinical context, which prevented their inclusion in the further analysis. As a result of this process, 27 publications remained and were subjected to a detailed full-text review. After a thorough evaluation of the full texts, publications containing the following criteria were excluded:

- Additional diseases ($n = 3$) – studies that involved conditions other than those specified in the inclusion criteria, such as neurological disorders unrelated to psychopathology.
- Genetic research ($n = 2$) – publications primarily focused on genetic research, which did not meet the inclusion criteria regarding imaging technologies and AI.

- Additional measurement methods ($n = 2$) – studies that applied measurement methods other than those specified in the inclusion criteria, which could have introduced ambiguity in the interpretation of results.

In the case of excluding 7 studies, three main reasons were particularly important in the context of the full-text analysis. The studies were excluded due to additional diseases ($n = 3$), which involved conditions other than those specified in the inclusion criteria, such as neurological disorders unrelated to psychopathology, including Parkinson's disease and Alzheimer's disease. These publications were excluded based on the criterion "Research on Parkinson's disease or Alzheimer's disease." Another reason was the focus on genetic studies ($n = 2$), which mainly concentrated on genetic aspects rather than imaging technology or artificial intelligence. In this case, the exclusion criterion applied was "Studies that combine genetic data." The final reason concerned the use of additional measurement methods ($n = 2$), which differed from those specified in the inclusion criteria, potentially leading to ambiguity in the interpretation of results. The exclusion of these studies was in line with the general methodological criteria.

Following this selection process, 27 publications remained and were included in the further systematic review. Ultimately, 20 articles were included in the review.

The correct procedure for creating a systematic review from systematic reviews is to include only those reviews that meet the quality criteria and are directly relevant to the topic under scrutiny rather than including all reviews published to date on the topic. Therefore, the number of obtained reviews is small and closely related to the developed PICOS scheme. Hence, within the scope of methodological assessment of the quality of systematic reviews found in PubMed, the tool: Assessment of Multiple Systematic Reviews (AMSTAR) was used. The AMSTAR tool is a questionnaire that assesses the presence of the following: prior design, duplicate selection and data extraction, comprehensive literature search, reference to the publication status as one of the inclusion criteria, presentation of the list of included/excluded publications, presentation of the characteristics of the included studies, documentation of the qualitative assessment of the studies included in the systematic review, appropriate formulation of conclusions, appropriate use of methods to synthesise the results, assessment of the likelihood of methodological errors of the included studies, documentation of conflicts of interest [18]. During the assessment on the AMSTAR scale, 11 questions were answered (yes, no, can't answer, not applicable). An affirmative answer (YES) received 1 point, a negative answer (No, can't answer) – 0, not applicable – 0. The quality of the review was determined on a scale of 1 to 11. Scores of less than five "yes" answers indicate low-quality reviews. Scores above or equal to five indicate a review of moderate quality. A score of ≥ 9 is a high quality review (Table 3).

The credibility assessment of nine systematic reviews based on the AMSTAR scale found four with moderate credibility and five with high credibility (Table 3). Papers [20, 25, 26, 36, 38, 40–42], and [7] received moderate scores, indicating minor deficits. The highest score of 11 points was awarded to [39], while [19] and [22] received 10 points, and [21, 23, 24, 37] received 9 points, indicating high credibility.

The first high-scoring review examined machine learning (ML) models used to diagnose bipolar disorder (BD), with SVMs, ANNs, ensemble models, linear regression, and

Table 3. AMSTAR Quality Assessment of the Systematic Reviews. The table presents the total number of “YES” answers for each systematic review based on the AMSTAR criteria. The quality assessment of each review is categorized as follows: 10 or more “YES” answers correspond to “High quality,” 8–9 “YES” answers indicate “Moderate quality,” and 5–7 “YES” answers represent “Low quality.”

Article	Total ‘YES’ answers	Quality Assessment
Jan Z, Ai-Ansari N “The Role of Machine Learning in Diagnosing Bipolar Disorder: Scoping Review”[19]	10	High quality
Victoria Welch. “Use of Mobile and Wearable Artificial Intelligence in Child and Adolescent Psychiatry: Scoping Review” [20]	8	Moderate quality
Mirjam Quaak “Deep learning applications for the classification of psychiatric disorders using neuroimaging data: Systematic review and meta-analysis” [21]	9	High quality
Marzena Oliveira Ribas Previously Marzena Szkodo “Technologies to support the diagnosis and/or treatment of neurodevelopmental disorders: A systematic review” [22]	10	High quality
Alban Lejeune “Use of Social Media Data to Diagnose and Monitor Psychotic Disorders: Systematic Review” [23]	9	High quality
Caio Pinheiro Santan “rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis” [24]	9	High quality
Matthew Bracher-Smith “Machine learning for genetic prediction of psychiatric disorders: a systematic review” [25]	6	Moderate quality
Eleonora Minissi “Assessment of the Autism Spectrum Disorder Based on Machine Learning and Social Visual Attention: A Systematic Review” [26]	8	Moderate quality
Barnaly Rashid “Towards a brain-based predictome of mental illness” [7]	6	Moderate quality
Guo Z, “Systematic review and meta-analysis: multimodal functional and anatomical neural alterations in autism spectrum disorder” [36]	8	Moderate quality
Di Camillo “Magnetic resonance imaging-based machine learning classification of schizophrenia spectrum disorders: a meta-analysis”[37]	9	High quality

(continued)

Table 3. (continued)

Article	Total ‘YES’ answers	Quality Assessment
Nickson D, “Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review” [38]	7	Moderate quality
Ding Y, “Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis” [39]	11	High quality
Chiu YH, Exploring the Role of Mobile Apps for Insomnia in Depression: Systematic Review” [40]	8	Moderate quality
Rahul J,” A systematic review of EEG based automated schizophrenia classification through machine learning and deep learning” [41]	5	Low quality
Wider W, „Mapping the evolution of neurofeedback research: a bibliometric analysis of trends and future directions” [42]	5	Low quality

Gaussian process models being the most common. The goal was to improve diagnostic speed and accuracy [39].

The second high-scoring review revealed that 68.8% of studies used technology for diagnosis rather than treatment of neurodevelopmental disorders (NDDs), with ML being the most common technology (61.8%). Neurofeedback was the most popular treatment technology. The review covered 221 studies, with 85.6% focusing on diagnosis, particularly autism and ADHD [21].

Lejeune et al. [23] reviewed AI’s potential to use social media data for diagnosing psychotic disorders, showing strong performance in identifying schizophrenia, but challenges remain in data quality. Santana et al. [24] analyzed machine learning classifiers for autism diagnosis, finding sensitivity and specificity rates of 73.8% and 74.8%, respectively, with SVM classifiers showing estimates above 76%.

Welch et al. [20] noted the lack of randomized controlled trials in AI research in child psychiatry. Nature [25] reported mixed performance of ML in predicting psychiatric disorders from genetics, with AUCs ranging from 0.48 to 0.95. A study by [26] suggested that ML could aid in early autism diagnosis through visual attention.

Recent research highlights AI’s growing role in mental disorder diagnosis. Guo et al. [36] showed AI’s effectiveness in differentiating autism spectrum disorder (ASD) patients from neurotypical individuals using neuroimaging. Di Camillo et al. [37] confirmed the potential of ML in classifying schizophrenia. Nickson et al. [38] found ML improved depression diagnosis using electronic health records. Ding et al. [39] showed that deep learning models are highly effective in diagnosing ASD, outperforming traditional methods. Chiu et al. [40] examined AI-powered mobile apps for managing insomnia and depression, demonstrating their increasing use in digital health. Rahul et al. [41]

highlighted AI's role in improving electroencephalography-based schizophrenia classification, while Wider et al. [42] found growing AI use in neurofeedback applications for mental health.

These studies demonstrate AI's transformative potential in enhancing diagnostic accuracy, efficiency, and early detection across psychiatric conditions.

As the AMSTAR scale is dedicated to the assessment of systematic reviews, it could not be used to assess the clinical trials found during the Cochrane Library search. The five-point Jadad scale was used to assess the quality of the clinical trials included in the review [27], the range of score quality: 0–2 Low, 3–5 High [28], Table 4.

Table 4. Assessment of Methodological Quality of Clinical Trials Using the Jadad Scale. The Jadad scale was used to assess the quality of the clinical trials included in this review. Scores range from 0 to 2 (Low quality) and 3 to 5 (High quality). The table presents the scores for key methodological criteria and the final quality rating of each study.

Clinical Trial	Study 1 [29]	Study 2 [30]	Study 3 [31]	Study 4 [32]
Randomized Study	1	1	1	1
Randomization Scheme	1	1	1	1
Double-blind	0	0	0	0
Blinding Method	0	0	0	0
Dropouts/Withdrawals	0	1	1	1
Final Score	2	3	3	3
Study Quality	Low	High	High	High

As shown, the results from the quality assessment of the clinical trials are generally satisfactory. Three of the studies were rated as high quality [30–32], while one was rated as low quality. The low rating is attributed to the preliminary nature of the study [29], which was still in the volunteer recruitment phase. As a result, the study lacked a description of dropouts and withdrawals, leading to the loss of a point in the evaluation. Despite this low score, the study was included in the review due to its relevance to the subject matter and its fulfillment of the inclusion and exclusion criteria for this systematic review. In all four studies, 0 points were assigned for questions 3 and 4, which is due to the nature of the clinical studies involved. These studies do not focus on drug therapies but rather on the use of artificial intelligence technologies for predicting, diagnosing, and treating mental illness. Therefore, it would be difficult to apply the double-blind method in these studies.

The study “Efficacy of the Mobile Application for Prediction and Prevention of Mood Episode Recurrence Based on Machine Learning” (CRM) was designed for a 12-month period, during which data were collected and compared between the active and placebo groups [29]. It involved male and female patients aged 19 to 70, diagnosed with Bipolar I Disorder, Bipolar II Disorder, or Major Depressive Disorder (MDD) based on DSM-5 criteria, who had been in a euthymic state for at least two weeks before recruitment. Participants were required to be Android smartphone users, able to install

and run the CRM application, and willing to wear a Fitbit device continuously, regularly synchronizing and backing up the collected data.

In Kannampallil et al. [30], machine learning algorithms were developed to predict depression remission in patients undergoing six months of problem-solving therapy (PST) using data from two randomized trials. The models, utilizing baseline characteristics and intervention engagement data, accurately predicted depression remission at both baseline and 2-month intervals in internal and external validation cohorts. These ML models offer potential for identifying likely responders to PST and optimizing early treatment for depression care.

In [31], the authors aimed to identify depressive symptom change patterns during psychotherapy for late-life depression with executive dysfunction. They also sought to predict nonresponse by mid-treatment and identify baseline characteristics associated with early nonresponse. The study, involving 221 older adults, found two distinct depression change trajectories through Latent Growth Mixture Models (LGMM), which were linked to treatment response at the study's conclusion. Random forest machine learning models identified significant modifiable predictors of persistent depression early in treatment, including low social support, high neuroticism, low treatment expectancy, and low perception of therapist acceptance. These findings suggest the potential for targeted interventions to address modifiable risk factors and improve therapy outcomes.

In Pratap et al. [32], the authors investigated the accuracy of passive phone sensors in predicting daily mood. This study demonstrated the potential of using smartphone-based sensors to track users' daily mood, based on data from passive sensors like GPS, accelerometers, and other mobile phone sensors. The research found that these sensors could offer valuable insight into daily mood patterns, although accuracy varied depending on individual behavioral patterns and sensor data features. This study adds to the growing body of research on utilizing mobile phone technologies to monitor and predict mood states, contributing to the broader field of digital mental health solutions.

Finally, the study presented in Table 4 explored the predictive ability of smartphone-based data for daily mood in a nationally recruited sample. While overall estimations revealed a weak association between physical mobility and daily mood, predictive models did not perform well for the entire sample (median $R^2 \sim 0$). However, personalized models combining various features showed improved predictions for individual mood changes, with GPS-derived mobility being a key contributor to mood prediction. This indicates limitations in using passive smartphone data for mood prediction at the population level due to significant variations in phone usage patterns and daily mood ratings. Nevertheless, there is potential for personalized models to predict individual mood states more accurately.

4 Results and Discussion

In this scoping review, we applied a comprehensive search strategy between January 2020 and January 2025. After searching the PubMed and Cochrane Library databases, we obtained 141 potential articles from PubMed and 37 trials from the Cochrane Library, as shown in Table 2. A multi-stage selection process was then applied to narrow down the articles to those most relevant to the research topic.

In the first stage, we reviewed the titles of 177 publications. Based on this preliminary selection, 84 articles were excluded due to lack of relevance to the research topic or non-compliance with the inclusion criteria, such as a focus on drug therapies rather than AI-based approaches to diagnosing, predicting, or treating mental health disorders. Following this, 93 publications advanced to the abstract review stage. After reviewing the abstracts, 66 articles were excluded, mainly due to their lack of relevant data on imaging technologies or AI, or an inappropriate clinical context for inclusion in the analysis. As a result, 27 publications proceeded to full-text review.

Ultimately, 20 articles met the inclusion criteria and were included in the final systematic review. The detailed selection process and the reasons for exclusions are presented in Fig. 1. Of note, no duplicates were found in the search results across the two databases.

The studies included in this review represent a diverse range of applications of AI and machine learning in the context of mental health, with a focus on predicting, diagnosing, and treating mood disorders.

One item regarding clinical trials is not listed in Table 5 because it has not been completed [29]. The aim of the study was to evaluate the effectiveness of a mobile application called Circadian Rhythm for Mood (CRM), which was developed to prevent the recurrence of mood disorders (major depression, bipolar disorder type 1 and 2) based on machine learning. Table 5 provides a summary of studies on machine learning in mental health disorders, and the study involving the CRM application will be included in future updates once completed.

Understanding the profound impact of precise diagnostics on mental health treatment, our review underscores the pivotal role of technology, particularly Machine Learning, in revolutionizing psychiatric diagnoses. Effective and accurate diagnosis of mental disorders is crucial for initiating and selecting effective treatments. The present review aimed to fill the gap in scoping reviews by gathering and categorizing ML models utilized in diagnosing psychiatric conditions. Growth has been observed recently in managing psychiatric disorders in child and adolescent psychiatry [20]. Studies have demonstrated how technologies can facilitate faster and improved diagnoses of various disorders, such as Bipolar Disorder, Major Depressive Disorder, Autism Spectrum Disorder [33], Attention-Deficit/Hyperactivity Disorder, Social Anxiety Disorder and schizophrenia [34, 35]. This trend aims to promote their utilization in clinical decision-making within the healthcare industry. Our review queries did not encompass the terms related to specific ML algorithms or models, resulting in the omission of articles using these terms in titles or abstracts instead of “ML,” thereby reducing the diversity of our scoping review. Nevertheless, our review is particularly relevant considering the increased interest in utilizing digital technologies. The ubiquitous nature of smartphone technology enables scalable approaches, such as treatment response or modification, at a fraction of the cost of clinic visits. Optimal clinical decision-making in treating depression should rely on more regular monitoring of symptoms and treatment outcomes rather than infrequent self-reports obtained during clinic visits. Research shows that mood states are best predicted at an individualized level, by examining one’s deviation from the norm rather than comparing it with a population norm [32]. However, such individualized approaches are relatively scarce and less prevalent in scientific publications. The role of AI in psychiatric neuroimaging has gained significant traction, particularly in Autism Spectrum

Table 5. Summary of Studies on Machine Learning in Mental Health Disorders.

Authors (Year)	Disorder(s)	Sample Size	Key Findings
Rashid et al. [7] (2020)	Multiple (SZ, MDD, BP, ASD, ADHD, etc.)	250 studies	Review of neuroimaging-based predictive approaches across various mental illnesses; highlights trends and shortcomings
Jan et al. [19] (2021)	Bipolar Disorder (BD)	15 to 25,000 participants	ML models used to diagnose chronic and first-episode BD, including SVM, ANNs, decision trees, and clustering algorithms
Welch et al. [20] (2022)	Autism Spectrum Disorder (ASD), ADHD, Anxiety	344 articles, 19 included	Focus on wearable biosensors (e.g., Fitbit) for passive data collection on physiological signals
Quaak et al. [21] (2021)	ASD, Schizophrenia, ADHD	ASD (22), SZ (22), ADHD (9)	Comparison of Deep Learning (DL) vs. Standard ML models in classifying psychiatric disorders, suggesting DL superiority
Szkodo et al. [22] (2023)	Autism Spectrum Disorder (ASD), Neurodevelopmental Disorders	4,108 records	Promising results for technology-based diagnosis and intervention in NDDs, including ASD
Lejeune et al. [23] (2022)	Schizophrenia	51 to 265,396 participants	Use of social media monitoring to track patients at risk of psychosis and detect schizophrenia relapses
Santana et al. [24] (2022)	Autism Spectrum Disorder (ASD)	55 studies	ML classifiers using Rs-fMRI data showed sensitivity and specificity around 75% for ASD diagnosis

(continued)

Table 5. (continued)

Authors (Year)	Disorder(s)	Sample Size	Key Findings
Bracher-Smith et al. [25] (2021)	Multiple Psychiatric Disorders (SZ, BD, OCD, etc.)	77 models across 13 studies	Performance of ML methods varied across disorders (e.g., AUC: SZ 0.54–0.95, BD 0.48–0.65, Autism 0.52–0.81)
Minissi et al. [26] (2022)	Autism Spectrum Disorder (ASD)	11 studies	ML algorithms used to distinguish children with ASD from typical developmental peers based on social attention
Kannampallil et al. [30] (2022)	Depression	Data from 2 randomized trials	ML models predicted depression remission for patients undergoing problem-solving therapy
Solomonov et al. [31] (2021)	Depression	221 older adults with major depression	Latent Growth Mixture Models (LGMM) predicted remission and response, highlighting factors like low treatment expectancy
Pratap et al. [32] (2019)	Depression	271 Android phone users	Daily mobility patterns from phone sensors correlated with depressive symptom severity
Guo et al. [36] (2024)	Autism Spectrum Disorder (ASD)	3,971 (ASD, TD participants)	ML and deep learning models used to analyze neuroimaging data to differentiate ASD patients from neurotypical individuals

(continued)

Table 5. (continued)

Authors (Year)	Disorder(s)	Sample Size	Key Findings
Di Camillo et al. [37] (2024)	Schizophrenia Spectrum Disorders (SSD)	12,723 SSD, 13,196 HCs	Deep learning models demonstrated high accuracy in distinguishing schizophrenia patients from healthy controls
Nickson et al. [38] (2023)	Depression	19 studies	ML applications for diagnosing depression using EHRs, improving diagnostic efficiency
Ding et al. [39] (2024)	Autism Spectrum Disorder (ASD)	9,495 patients	Review of deep learning approaches for ASD diagnosis, confirming CNNs outperform traditional methods in sensitivity
Chiu et al. [40] (2024)	Insomnia & Depression	18 studies	AI-powered mobile apps for insomnia and depression management, providing personalized treatment recommendations
Rahul et al. [41] (2024)	Schizophrenia	12 studies	EEG-based schizophrenia classification using ML, enhancing classification accuracy for early detection
Wider et al. [42] (2024)	Neurofeedback Research	3,626 articles	Bibliometric analysis of AI applications in neurofeedback, identifying trends and future directions for interventions

Disorder (ASD), as highlighted by Guo et al. (2024) [36]. Their systematic review and meta-analysis demonstrate how AI-driven neuroimaging analyses effectively differentiate ASD patients from neurotypical individuals, reinforcing the potential of ML in improving diagnostic precision. These findings emphasize the importance of integrating multimodal neuroimaging techniques to unravel the complex neural alterations associated with ASD, fostering advancements in early and accurate detection. The utilization

of platforms and smartphones in mental health treatment is rapidly evolving. These devices offer myriads of possibilities, from health monitoring to providing interactive therapeutic tools. Smartphones, almost ubiquitous in everyday life, serve as a key gateway to mental health-related services. Mobile applications allow users to monitor mood, sleep, physical activity and daily routines. The tracking enables patients to keep tabs on their well-being, providing valuable reference points for medical professionals. These platforms also offer educational content, counseling, and even therapeutic interventions such as mindfulness training or relaxation tools. By integrating game elements with mental health-related tasks, these applications encourage consistent engagement, a crucial aspect of the therapeutic process. Collaboration with mental health professionals via telemedicine platforms is increasingly popular. Online therapy sessions, mental health advice, and the ability to connect with therapists through smartphones are reshaping the way we perceive and engage with mental health treatment. However, for these modern technologies to be fully effective, it is imperative to ensure patient data privacy and security. With the growing use of these platforms in mental healthcare, continuous refinement of IT solutions and regulatory frameworks becomes a priority to ensure effective and secure tools in mental health therapy.

However, we must take into account platform heterogeneity, significant differences between iOS and Android platforms, impact passive data functions, detail, and sampling frequency. For instance, iOS limits access to call logs and messages. Lack of adherence to appropriate procedures for machine learning is also a common reason why a model is assessed as at high risk of bias. Smartphones offer remarkable avenues for scalable approaches and platform heterogeneity; inconsistencies pose challenges in passive data collection, which results in the need for standardized protocols for uniformity and reliability. Standard model validation procedures were applied, yet by not all researchers [25]. Another growing domain in AI-driven mental health research is neurofeedback, as evidenced by the bibliometric analysis conducted by Wider et al. (2024) [42]. Their findings highlight the increasing integration of AI in neurofeedback applications, identifying key research trends and future directions for AI-enhanced interventions. By leveraging real-time neurofeedback mechanisms, AI-based models can potentially enhance therapeutic outcomes, paving the way for more personalized and adaptive treatment strategies in mental healthcare.

Furthermore, acknowledging the potential of AI in neuroimaging for developing diagnostic models, particularly the promising strides made by Deep Learning in psychiatric neuroimaging, heralds a new era of diagnostic accuracy and efficacy [21].

Our analysis indicates that AI and machine learning ML are promising tools for diagnosing and predicting psychiatric disorders. However, their integration into clinical practice still faces significant challenges. One key finding from the reviewed studies is the superiority of deep learning methods over traditional ML algorithms in neuroimaging analysis (MRI, fMRI, EEG). Neural network-based models, particularly CNNs, demonstrated higher accuracy in classifying patients with autism spectrum disorder ASD and schizophrenia compared to conventional techniques.

A noticeable shift in research trends has also been observed. While early studies primarily focused on AI for diagnostic purposes, recent research is increasingly exploring AI's role in predicting treatment response and personalizing therapy. AI models that

analyze neuroimaging data, combined with mobile sensor data (e.g., smartphone-based tracking, wearable biosensors), are being tested as decision-support tools in clinical settings.

Despite promising results, several limitations and barriers must be addressed before AI can be widely adopted in psychiatric practice:

- Data heterogeneity – The lack of standardized protocols for data collection makes it difficult to compare study results and validate AI models across different clinical populations.
- Validation standards – Many studies rely on small, non-representative datasets, limiting the generalizability of findings. Moreover, few studies incorporate randomized controlled trials (RCTs), which are essential for evaluating the effectiveness of diagnostic AI applications.
- Ethical and legal concerns – Issues related to patient privacy, algorithmic bias, and the risk of misdiagnosis without human oversight remain significant challenges for AI implementation in psychiatry.

To facilitate the effective integration of AI in psychiatry, future research should focus on improving data quality, increasing model transparency, and developing hybrid approaches that combine neuroimaging with behavioral and genetic data. Such strategies could enhance diagnostic precision and treatment personalization. Additionally, establishing standardized validation frameworks and ensuring ethical AI deployment in clinical settings will be crucial for bridging the gap between research and real-world applications.

5 Conclusion

This review highlights the evolving role of artificial intelligence in psychiatric diagnostics, with a particular focus on neuroimaging-based classification and treatment prediction. Deep learning models, especially convolutional neural networks, demonstrate superior accuracy in identifying psychiatric conditions such as autism spectrum disorder and schizophrenia. Additionally, AI is increasingly being used not only for diagnosis but also for predicting treatment response, particularly in mood disorders, by integrating neuroimaging with behavioral and physiological data.

A key finding of this review is the emerging shift toward AI-driven personalized psychiatry, where machine learning models analyze patient-specific patterns rather than relying solely on population-based norms. However, significant challenges remain, including data heterogeneity, lack of standardized validation frameworks, and ethical concerns related to privacy and algorithmic bias. While standardization is essential for ensuring model reliability, it must be balanced with the need for adaptive, patient-specific AI applications that support individualized treatment strategies.

Future research should focus on developing standardized yet flexible AI frameworks that enhance diagnostic accuracy while allowing for personalization. Bridging this gap between AI-driven insights and clinical applicability will be crucial for integrating AI into psychiatric practice in a reliable, ethical, and effective manner.

Declarations of Interest. None.

References

1. Gong B et al (2019) Neuroimaging in psychiatric disorders: a bibliometric analysis of the 100 most highly cited articles. *J Neuroimaging* 29:14–33. <https://doi.org/10.1111/jon.12570>
2. Najafpour Z, Fatemi A, Goudarzi Z, Goudarzi R, Shayanfar K, Noorizadeh F (2021) Cost-effectiveness of neuroimaging technologies in management of psychiatric and insomnia disorders: a meta-analysis and prospective cost analysis. *J Neuroradiol* 48:348–358. <https://doi.org/10.1016/j.neurad.2020.12.003>
3. Waller J, Miao T, Ikedionwu I, Lin KK (2021) Reviewing applications of structural and functional MRI for bipolar disorder. *Jpn J Radiol* 39:414–423. <https://doi.org/10.1007/s11604-020-01074-5>
4. Iorio-Morin C, Sarica C, Elias GJB, Harmsen I, Hodaie M (2022) Neuroimaging of Psychiatric Disorders. In: *Progress in Brain Research*, Elsevier, Vol 270, pp 149–169 ISBN 978-0-12-823771-7
5. Paul SM, Potter WZ (2023) Finding New and Better Treatments for Psychiatric Disorders. *Neuropsychopharmacol*. <https://doi.org/10.1038/s41386-023-01690-5>
6. Mitelman SA (2019) Transdiagnostic neuroimaging in psychiatry: a review. *Psychiatry Res* 277:23–38. <https://doi.org/10.1016/j.psychres.2019.01.026>
7. Rashid B, Calhoun V (2020) Towards a brain-based predictive model of mental illness. *Hum Brain Mapp* 41:3468–3535. <https://doi.org/10.1002/hbm.25013>
8. Lin E, Lin C-H, Lane H-Y (2020) Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. *IJMS* 21:969. <https://doi.org/10.3390/ijms21030969>
9. Giansanti D (2023) Precision medicine 2.0: how digital health and AI are changing the game. *JPM* 13:1057. <https://doi.org/10.3390/jpm13071057>
10. Karpel I, Starcevic A, Urzeniczok M (2022) Database and AI diagnostic tools improve understanding of lung damage, correlation of pulmonary disease and brain damage in COVID-19. *Sensors* 22:6312. <https://doi.org/10.3390/s22166312>
11. Starcevic A, Vucinic B, Karpel I (2023) Training set preparation for deep model learning inpatients with ischemic brain lesions and gender identity disorder. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada JM, (eds.) *Artificial Intelligence and Soft Computing, Lecture Notes in Computer Science*; Springer Nature Switzerland: Cham, Vol 14126, pp 176–191 ISBN 978-3-031-42507-3
12. APA PsycArticles. <https://www.apa.org/pubs/databases/psycarticles>. Accessed 2 Nov 2023
13. Gómez Penedo JM, et al (2023) Using machine learning algorithms to predict the effects of change processes in psychotherapy: toward process-level treatment personalization. *Psychotherapy*. <https://doi.org/10.1037/pst0000507>
14. Sterner P, Goretzko D, Pargent F (2023) Everything has its price: foundations of cost-sensitive machine learning and its application in psychology. *Psychological Methods*. <https://doi.org/10.1037/met0000586>
15. Götz FM, Maertens R, Loomba S, Van Der Linden S (2023) Let the algorithm speak: how to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. <https://doi.org/10.1037/met0000540>
16. Tan R, Shoenberger H, Cong W, Mahdavi M (2023) Understanding the structural components behind the psychological effects of autonomous sensory meridian response (ASMR) with machine learning and experimental methods. *J Media Psychol* 35:181–189. <https://doi.org/10.1027/1864-1105/a000359>

17. Cochrane Handbook for Systematic Reviews of Interventions|Cochrane Training Available online: <https://training.cochrane.org/handbook>. Accessed 2 Nov 2023
18. Shea BJ et al (2007) Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 7:10. <https://doi.org/10.1186/1471-2288-7-10>
19. Jan Z, et al (2021) the role of machine learning in diagnosing bipolar disorder: scoping review. *J Med Internet Res* 23:e29749. <https://doi.org/10.2196/29749>
20. Welch V et al (2022) Use of mobile and wearable artificial intelligence in child and adolescent psychiatry: scoping review. *J Med Internet Res* 24:e33560. <https://doi.org/10.2196/33560>
21. Quaak M, Van De Mortel L, Thomas RM, Van Wingen G (2021) Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *NeuroImage: Clinical* 30:102584. <https://doi.org/10.1016/j.nicl.2021.102584>
22. (Previously Marzena Szkodo) MOR, Micai M, Caruso A, Fulceri F, Fazio M, Scattoni ML (2023) Technologies to support the diagnosis and/or treatment of neurodevelopmental disorders: a systematic review. *Neurosci Biobehav Rev* 145:105021. <https://doi.org/10.1016/j.neubiorev.2022.105021>
23. Lejeune A, Robaglia B-M, Walter M, Berrouiguet S, Lemey C (2022) Use of social media data to diagnose and monitor psychotic disorders: systematic review. *J Med Internet Res* 24:e36986. <https://doi.org/10.2196/36986>
24. Santana CP, De Carvalho EA, Rodrigues ID, Bastos GS, De Souza AD, De Brito LL (2022) Rs-fMRI and machine learning for ASD diagnosis: a systematic review and meta-analysis. *Sci Rep* 12:6030. <https://doi.org/10.1038/s41598-022-09821-6>
25. Bracher-Smith M, Crawford K, Escott-Price V (2021) Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol Psychiatry* 26:70–79. <https://doi.org/10.1038/s41380-020-0825-2>
26. Minissi ME, Chicchi Giglioli IA, Mantovani F, Alcañiz Raya M (2022) Assessment of the autism spectrum disorder based on machine learning and social visual attention: a systematic review. *J Autism Dev Disord* 52:2187–2202. <https://doi.org/10.1007/s10803-021-05106-5>
27. JADAD Scores - RCEMLearning India. <https://www.rcemlearning.org/modules/critical-appraisal-appraising-a-treatment-early-goal-directed-therapy/lessons/methodology-jadad-scores/topic/jadad-scores/>. Accessed 2 Nov 2023
28. On Biostatistics and Clinical Trials: Jadad Scale to Assess the Quality of Clinical Trials. <https://onbiostatistics.blogspot.com/2009/03/jadad-scale-to-assess-quality-of.html>. Accessed 2 Nov 2023
29. Efficacy of the Mobile Application for Prediction and Prevention of Mood... <https://trialbulletin.com/lib/entry/ct-05400785>. Accessed 17 Oct 2023
30. Kannampallil T et al (2022) Cross-Trial prediction of depression remission using problem-solving therapy: a machine learning approach. *J Affect Disord* 308:89–97. <https://doi.org/10.1016/j.jad.2022.04.015>
31. Solomonov N et al (2021) Modifiable predictors of nonresponse to psychotherapies for late-life depression with executive dysfunction: a machine learning approach. *Mol Psychiatry* 26:5190–5198. <https://doi.org/10.1038/s41380-020-0836-z>
32. Pratap A et al (2019) The accuracy of passive phone sensors in predicting daily mood. *Depress Anxiety* 36:72–81. <https://doi.org/10.1002/da.22822>
33. Ghafouri-Fard S, Taheri M, Omrani MD, Daaee A, Mohammad-Rahimi H, Kazazi H (2019) Application of single-nucleotide polymorphisms in the diagnosis of autism spectrum disorders: a preliminary study with artificial neural networks. *J Mol Neurosci* 68:515–521. <https://doi.org/10.1007/s12031-019-01311-1>
34. Vivian-Griffiths T et al (2019) Predictive modeling of schizophrenia from genomic data: comparison of polygenic risk score with kernel support vector machines approach. *Am J Med Genetics Pt B* 180:80–85. <https://doi.org/10.1002/ajmg.b.32705>

35. Trakadis YJ, Sardaar S, Chen A, Fulginiti V, Krishnan A (2019) Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes. *Am J Med Genetics Pt B* 180:103–112. <https://doi.org/10.1002/ajmg.b.32638>
36. Guo Z et al (2024) Systematic review and meta-analysis: multimodal functional and anatomical neural alterations in autism spectrum disorder. *Mol Autism* 15(1):16. <https://doi.org/10.1186/s13229-024-00593-6>. PMID:38576034;PMCID:PMC10996269
37. Di Camillo F, et al (2024) Magnetic resonance imaging-based machine learning classification of schizophrenia spectrum disorders: a meta-analysis. *Psychiatry Clin Neurosci* 78(12):732–743. <https://doi.org/10.1111/pcn.13736>. Epub 2024 Sep 18. PMID: 39290174; PMCID: PMC11612547
38. Nickson D, Meyer C, Walasek L, Toro C (2023) Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review. *BMC Med Inform Decis Mak* 23(1):271. <https://doi.org/10.1186/s12911-023-02341-x>. PMID:38012655; PMCID:PMC10680172
39. Ding Y, Zhang H, Qiu T (2024) Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis. *BMC Psychiatry* 24(1):739. <https://doi.org/10.1186/s12888-024-06116-0>. PMID:39468522;PMCID:PMC11520796
40. Chiu YH, Lee YF, Lin HL, Cheng LC (2024) Exploring the role of mobile apps for insomnia in depression: systematic review. *J Med Internet Res* 18(26):e51110. <https://doi.org/10.2196/51110>. PMID:39423009;PMCID:PMC11530740
41. Rahul J, Sharma D, Sharma LD, Nanda U, Sarkar AK (2024) A systematic review of EEG based automated schizophrenia classification through machine learning and deep learning. *Front Hum Neurosci* 14(18):1347082. <https://doi.org/10.3389/fnhum.2024.1347082>. PMID: 38419961;PMCID:PMC10899326
42. Wider W et al (2024) Mapping the evolution of neurofeedback research: a bibliometric analysis of trends and future directions. *Front Hum Neurosci* 10(18):1339444. <https://doi.org/10.3389/fnhum.2024.1339444>. PMID:38799297;PMCID:PMC11116792

Biomaterials, Nanomaterials, Regenerative Medicine, and Prosthetics



Using Backscattered Electrons to Determine Surface Charge on Laser-Treated Titanium Alloys

Arturs Abolins¹ , Alberta Aversa² , Marks Gorohovs³ , Yuri Dekhtyar³ ,
Maris Dortiņš³, Lyubomir Lazov⁴ , Arturs Mamajevs³, Hermanis Sorokins³ ,
and Edmunds Sprudz⁴

¹ Food Processing Laboratory, Rezekne Academy of Technologies, Pilsīela 23a,
Rezekne, Latvia

arturs.abolins@rta.lv

² Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli
Abruzzi, 24, Torino, Italy

alberta.aversa@polito.it

³ Riga Technical University, Mechanical and Biomedical Engineering Institute, Kipsalas 6B,
Riga, Latvia

{marks.gorhovs, jurijs.dehtjars, hermanis.sorokins}@rtu.lv

⁴ Rezekne Academy of Technologies, Atbrīvošanas Aleja 115, Rezekne, Latvia

lyubomir.lazov@rta.lv

Abstract. Oxidized titanium alloys, such as Ti 6Al 4V, are widely used in medical implants because their biocompatibility depends on surface properties like morphology, chemistry, and electrical charge. Laser processing in atmospheric, oxygenated conditions produces a thin oxide film with an embedded electrical charge, altering these surface characteristics. Although SEM is routinely employed to assess morphology and chemistry, its potential for evaluating surface charge is underexplored. In this study, we laser processed Ti 6Al 4V samples and used SEM in backscattered electron (BSE) mode to quantify surface charging. We correlated changes in BSE signal strength with work function measurements, finding that a work function increase from approximately 4.85 to 5.00 eV—indicative of greater surface charging—was accompanied by a measurable decrease in BSE image grey levels. Measurements at 10, 15, and 20 keV revealed a significant difference between 10 and 15 keV, while at 20 keV, the effect of surface charge on electron scattering was minimal. This approach demonstrates that BSE imaging can integrate charge analysis into routine SEM evaluations of laser-processed implant surfaces.

Keywords: laser processing · titanium oxide · surface charge · scanning electron microscopy

1 Introduction

Titanium alloys, particularly Ti-6Al-4V, are widely used in orthopedic and dental implants as well as coronary stents [1–4] because of their high biocompatibility, corrosion resistance, and favorable strength-to-weight ratio. Implant performance depends on effective cell attachment and osseointegration [5], both of which are influenced by surface properties such as morphology, chemistry, and electrical charge [6].

Laser processing is an effective method to functionalize implant surfaces by modifying these properties [6, 7], with the modifications relying on processing parameters such as power, pulse frequency, and scanning speed [8]. When a titanium surface is laser processed in atmospheric, oxygenated conditions, oxidation occurs to form a thin oxide layer with an embedded electrical charge [7, 9, 10]. The embedded charge arises primarily through dipole polarization and the formation of surface and bulk defects (e.g., oxygen vacancies and titanium interstitials) [8]. The altered surface charge can influence cell attachment and proliferation by modulating interfacial interactions [11].

Evaluating the different aspects of surface functionalization—morphology, chemistry, and charge—typically requires multiple instruments. Scanning electron microscopy (SEM) combined with energy dispersive spectroscopy (EDS) is routinely used to assess morphology and chemical composition, while near-threshold photoelectron emission spectroscopy (PES) [12] is employed to quantify surface charge via shifts in the photon energy required for photoelectron emission. This multi-instrument approach can be time-consuming and resource intensive.

Backscattered electron (BSE) imaging, an inherent SEM mode, is commonly used for compositional and topographical imaging because the intensity of backscattered electrons depends on atomic number and density [13]. However, the underlying physics also renders BSE imaging sensitive to local electric fields. Electron scattering is governed not only by atomic structure but also by the near-surface electrical field generated by embedded surface charge. At high incident electron energies (tens of keV), electrons have sufficient momentum to largely overcome the influence of these fields, resulting in stable grey levels in the BSE image [14]. At lower incident energies (e.g., 10 and 15 keV), however, the near-surface electrical field more strongly deflects electrons, leading to increased scattering and a reduction in image grey level. This physical phenomenon provides a basis for using BSE imaging to evaluate surface charge.

The aim of the present study is to demonstrate that BSE imaging can be incorporated as an additional tool in routine SEM evaluations to assess surface charge. Titanium oxide, with its critical role in cell attachment and implant biocompatibility, serves as a convenient model system for illustrating this integrated approach to surface characterization.

2 Materials

Titanium alloy (Ti-6Al-4V) samples, measuring $10 \times 10 \text{ mm}^2$, were prepared by cutting stock material supplied by Taizhou Yutai Metal Materials Co., China, using a Suntop ST-FC1390 fiber laser cutting machine (Suzhou, China) under ambient atmospheric conditions. After cutting, the samples were rinsed with technical-grade ethanol, and the specimen surfaces were immediately treated with a laser beam.

3 Methods

3.1 Laser Surface Processing

Surface processing was conducted using a Rofin PowerLine F20 Varia fiber laser (Rofin-Sinar, Bergkirchen, Germany) operating at a wavelength of 1064 nm. The maximum available power of the laser is 20 W, and the processing parameters were set as follows: a pulse frequency of 700 kHz and a pulse width of 4 ns. The laser beam was scanned across the specimen surface at a speed of 900 mm/s, with an effective treated width of 10 μm . A bidirectional scanning pattern was employed, alternating between left-to-right and right-to-left directions; scan lines were arranged at a 90° angle relative to each other with no overlap between adjacent lines. Each designated area, sized 5 $\times$ 5 mm², was subjected to a single scan, and four such areas were processed on each sample.

The treatment was performed under normal atmospheric conditions. Three power settings were applied to the four fields: one field processed at 63% Pmax (red), one field at 73% Pmax (orange), and two fields at 83% Pmax (green). The color changes were used to infer the formation of oxide layers with approximate thicknesses of 700 nm (red), 600 nm (orange), and 550 nm (green) [15].

After processing, the samples were wrapped in aluminum foil and stored in a plastic zip lock bag at ambient atmospheric conditions for three days before further measurements.

3.2 Surface Characterization

Surface charging was identified using near-threshold photoelectron emission spectroscopy (PES), which measures the emission of electrons escaping from the surface layer. In parallel, scanning electron microscopy (SEM) was performed in secondary electron (SE) mode to evaluate the sample surface morphology, in energy dispersive spectroscopy (EDS) mode to evaluate the sample surface chemical composition, and in backscattered electron (BSE) mode to evaluate the influence of the embedded charge on electron scattering.

3.2.1 Photoelectron Emission Spectroscopy (PES)

To induce photoelectron emission, the specimen is irradiated with ultraviolet (UV) light, causing photoelectrons to be emitted from a surface layer with an effective sampling depth of approximately 10–100 nm [12]. Photoelectron emission was detected using a custom-built spectrometer, with all measurements conducted in a vacuum maintained at $\sim 10^{-5}$ Torr throughout the experiment.

Electron emission was excited by UV photons with energies ranging from 4.2 to 6.2 eV, supplied by a 30 W deuterium lamp (model L7296, Hamamatsu Photonics, Japan). (Note: the lamp was not calibrated, and its stability was not measured.) A UV monochromator with automatic scanning (MDR-2, Russia) continuously selected wavelengths. Pulses were collected for one second at 1 nm increments as the monochromator swept from 300 nm to 200 nm over a 100-s period; the wavelengths were then converted into corresponding incident photon energies ($h\nu$).

The photoelectron emission current, I , was recorded at each photon energy, and the spectral dependence $I(h\nu)$ was constructed. The electron work function, ϕ , was determined by identifying the photon energy at which $I(h\nu)$ reaches zero—just before the signal rises above the noise level—expressed as $\phi = h\nu$ at $I(h\nu) = 0$. An increase in ϕ is interpreted as an indication of an increased negative surface charge [12].

To refine the estimation of ϕ , the relationship

$$I(h\nu) = a(h\nu - \phi)^m \quad (1)$$

was applied, where a is a proportionality coefficient and m is the power index (typically ranging from 1 to 3, depending on the electron transition mechanism) [16]. Although Eq. (1) is nonlinear, it can be linearized when $h\nu$ is close to ϕ . In practice, the linearized form

$$I(h\nu) = Ah\nu + B \quad (2)$$

was used for regression analysis within the tolerance defined by the uncertainty in I (given by the Poisson standard deviation, $I/2$) [17]. Under the condition $I(h\nu = \phi) = 0$, the work function was estimated as $\phi = B/A$. The relative uncertainty in ϕ , $\delta(\phi)$, was calculated from the relative uncertainties in A and B according to $\delta(\phi) = \delta(A) + \delta(B)$. The experiment demonstrated a relative uncertainty in ϕ of approximately 4%.

3.2.2 Scanning Electron Microscopy (SEM)

The processed surfaces were analyzed using a TESCAN S9000 scanning electron microscope (TESCAN, Brno, Czech Republic) in both Secondary Electron (SE) and Backscattered Electron (BSE) modes. The same SEM was also used to assess the chemical composition of the surfaces via Energy-dispersive X-ray spectroscopy (EDS). SEM operating conditions included a magnification of $\times 5,000$, a field of view of $56.6 \mu\text{m}$, and a working distance of 5 mm . The scan speed was set to 4, and the probe current was maintained at 100 pA . The electron beam energy was set to 10 keV . SE imaging provided topographical details, while BSE imaging was used to assess surface charging on the processed areas.

To be significantly influenced by the electrical charge of the specimen surface, the energy of the scattered electrons must be as low as possible. Because of the technical capabilities of the TESCAN S9000, the electron beam energy was varied between 10 keV and 20 keV . At these energies, backscattered electrons are emitted from a depth of approximately $10\text{--}100 \text{ nm}$ [], which is comparable to the thickness of the oxide layer produced by laser processing. This correspondence ensures that the detected BSE signal is representative of the laser-fabricated oxide.

Surface charging was evaluated by measuring the signal produced by the backscattered-electron (BSE) detector at various tilt angles against the plane normal to the e-beam. The stage was configured to vary the specimen's orientation from $+40^\circ$ to -40° relative to the electron beam's normal incidence, in 10° increments. During each tilt setting, the BSE signal was recorded using the BSE detector positioned within the optical column of the SEM. By changing the tilt angle it is possible to identify how surface charging—affected by the oxide layer and local conductivity—modified the emission and detection of backscattered electrons, allowing for a direct comparison



Fig. 1. Scheme of BSE detection

of tilt-dependent BSE responses across different sample regions. A graphical representation of the setup can be seen in Fig. 1. These measurements were performed with an incident electron beam energy of 10 keV and with a scan length of 5 mm in perpendicular and symmetrically to the border between the neighboring colored areas (“red-orange” and “green-green”). At the small angle between the incident beam and the BSE measured surface its electrical charge had the strongest impact to scatter reflected electrons. In this way, the electrical field by the titanium oxide charge could be identified (Fig. 2).

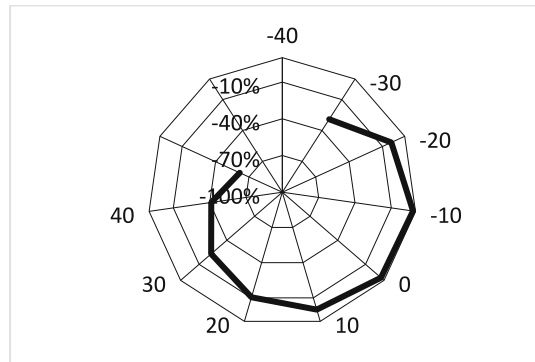


Fig. 2. Relative angular increment of the BSE current for laser treatment at 73% Pmax mode at the 10 keV incident e-beam.

4 Results and Discussion

Figure 3 presents the results of EDS measurements performed on surfaces processed at different laser power settings (%Pmax).

A distinct peak near 0.28 keV is observed in all samples, indicating the presence of carbon—most likely adventitious carbon—with comparable intensity across the surfaces. The titanium $L\alpha$ emission peak appears around 0.45 keV. The oxygen $K\alpha$ emission

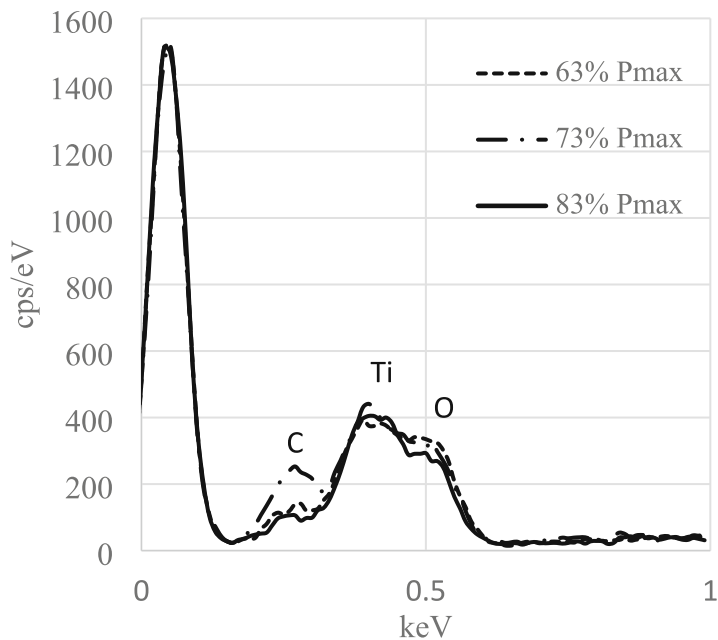


Fig. 3. EDS spectra of the alloy specimens’ surfaces processed at different laser power settings.

peak, observed near 0.52 keV, is less intense in the 83% Pmax sample compared to the 63% and 73% Pmax samples, suggesting a lower oxygen concentration. Adjacent to this, a peak near 0.51 keV, attributed to vanadium, shows a similar relative behavior. Overall, these results confirm that all laser-treated surfaces possess the expected base composition of Ti 6Al 4V and exhibit the formation of a titanium oxide layer.

Figure 4 shows SEM and BSE images of the treated sample surfaces. The SEM images reveal a distinctive “mud-crack” pattern in the oxide layer—likely resulting from thermal expansion mismatches and rapid cooling that fracture the film into polygonal regions. Specifically, the 63% Pmax surface displays tightly interlocked cracks with sharply defined ridges and occasional blister-like spots; the 73% Pmax surface shows broader, smoother crack edges with fewer nodules; and the 83% Pmax surface appears more irregular, with prominent nodular or blistered areas and rougher crack edges. In the corresponding BSE images, the same crack network is visible as dark lines; however, variations in grey level emerge within the individual “tiles.” For the 63% Pmax surface, the regions between cracks are mostly uniform in grey level, aside from a few darker spots that may indicate thinner oxide regions. The 73% Pmax surface exhibits uniform brightness with only subtle tonal gradients near cracks, whereas the 83% Pmax surface shows nodular features with variable brightness, suggestive of local variations in surface charge.

Figure 5 presents the increments in BSE grey level as a function of the specimen tilt angle (α) for different %Pmax settings. Here, the grey level increment is defined as the difference between the BSE grey level at a given tilt angle and the baseline grey level at 0° tilt. The angular dependence of BSE intensity—arising from changes in the relative

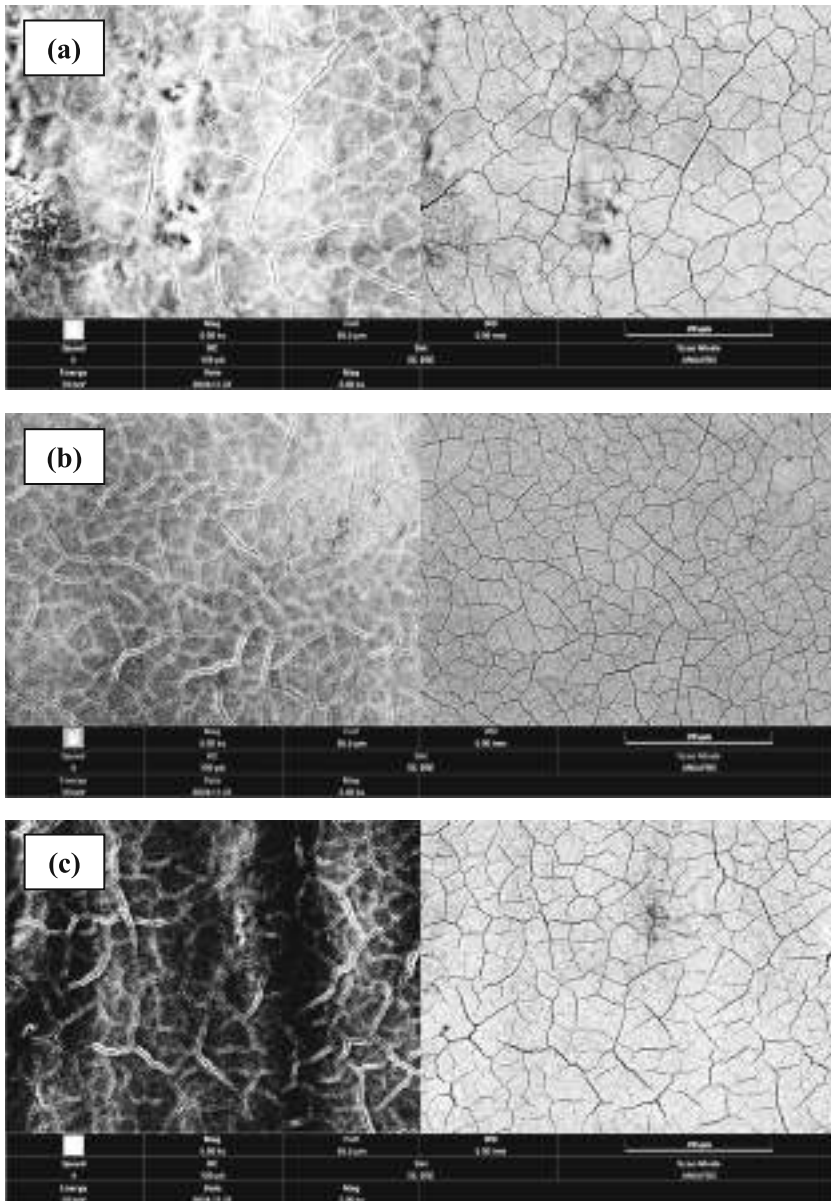


Fig. 4. SE (left) and BSE (right) images of the Ti -oxide layers laser treated at different laser power values: a) 63% $P_{\max}$, b) 73% $P_{\max}$, c) 83% $P_{\max}$

angles between the specimen surface, the incident electron beam, and the detector—is well known [9]. The results in Fig. 5 agree with this behavior; however, the curves obtained at an incident electron beam energy of 10 keV significantly deviate from those measured at 15 keV and 20 keV, especially for tilt angles greater than 10° . This stronger

deviation at lower energy and larger tilt angles supports the hypothesis that the surface electrical field is influencing electron scattering.

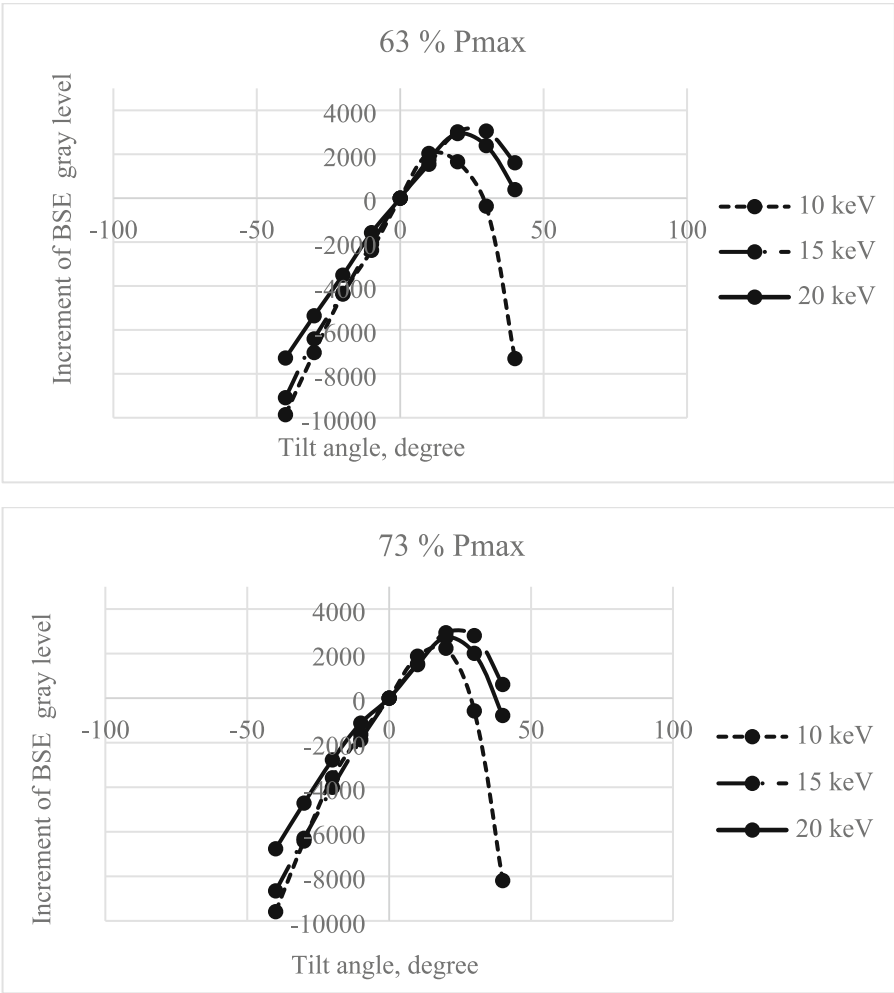


Fig. 5. Increments of the BSE grey level in dependence on the tilt angle for different % $P_{\max}$.

Figure 6 illustrates the influence of % $P_{\max}$ on the BSE grey level increment at a tilt angle of 40° , where the deviation is most pronounced. Since the BSE grey level is mainly affected by low-energy electrons, the observed decrease in grey level with increasing % $P_{\max}$ reinforces the presence of a surface electrical field. In other words, as % $P_{\max}$ increases, the negative surface charge becomes more pronounced, which is reflected in the reduced BSE grey level and is consistent with a corresponding rise in the electron work function.

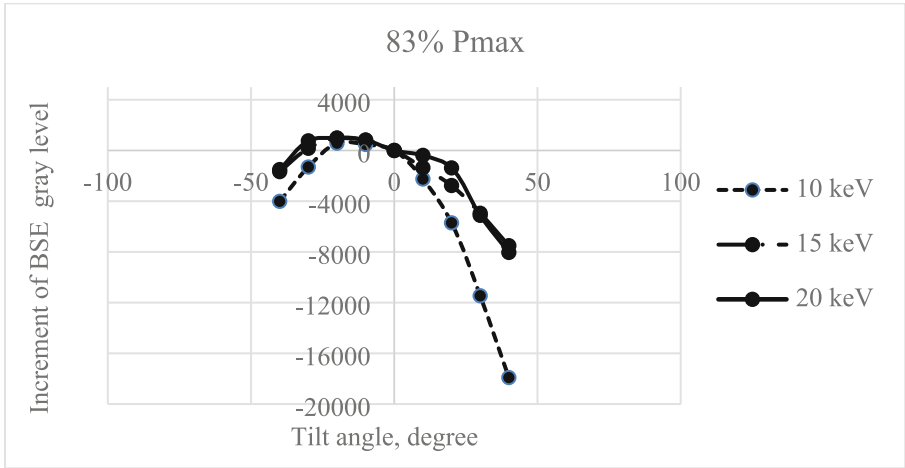


Fig. 5. (continued)

The reproducibility of the measured digital grey level (DGL) was found to be greater than 93%, confirming the consistency of the BSE measurements across repeated trials.

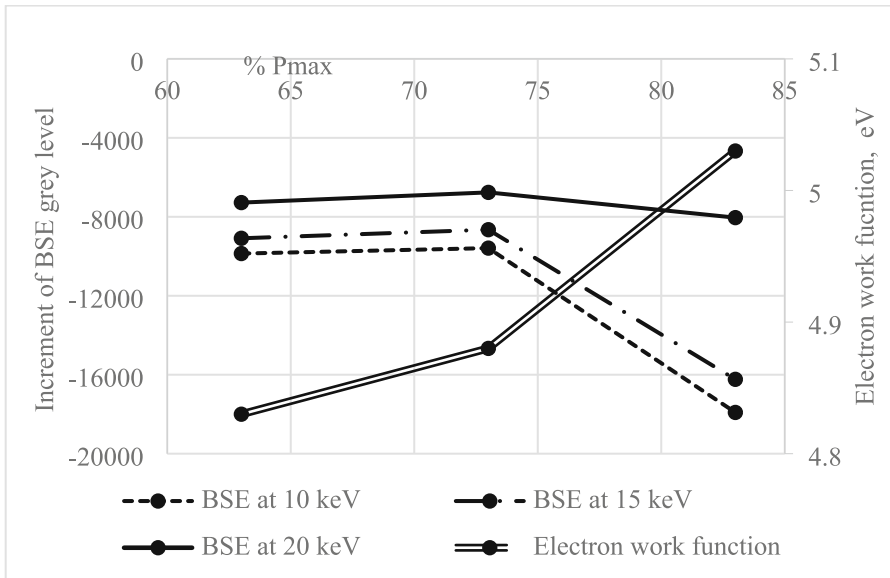


Fig. 6. Influence of $\%P_{max}$ on the incremented of the BSE grey level identified at the tilt angle of 40 degrees.

5 Conclusion

The angular distribution of backscattered electrons, as measured by BSE imaging, reveals a deflection pattern attributable to the near-surface electrical field. When these measurements are conducted at low incident electron beam energies (around 10 keV or lower) and with the detector positioned at a minimal angle relative to the specimen surface, the sensitivity to surface charge is maximized. This BSE approach, in conjunction with PES data, supports the determination of the surface electrical charge induced by laser processing.

Under the specified laser treatment conditions—1064 nm wavelength, 63–83% of the maximum 20 W power, 700 kHz pulse frequency, 4 ns pulse width, and 900 mm/s scanning speed—the Ti-6Al-4V alloy surface consistently acquires a negative electrical charge. This conclusion is supported by PES measurements, which show an increase in the electron work function, and by BSE measurements that reveal a corresponding decrease in grey level intensity at low electron energies. Furthermore, the magnitude of the negative charge increases with higher laser power settings.

Acknowledgement. The research was supported by the project Electrical Potential and Cell Immobilization Capacity of Laser-Processed Titanium Alloy Surfaces (Nr. RTU-PA-2024/I-0051) funded from the European Union Recovery and Resilience Facility project Nr. 5.2.1.1.i.0/2/24/I/CFLA/003.

References

1. Oldani C, Dominguez A (2012) Titanium as a biomaterial for implants. *Recent Adv Arthroplasty* 218:149–162. <https://doi.org/10.5772/27413>
2. Wang W, Poh CK (2013) Titanium alloys in orthopaedics. *Titanium Alloys-Adv Prop Control* 15:1–20. <https://doi.org/10.5772/55353>
3. Nicholson JW (2020) Titanium alloys for dental implants: a review. *Prosthesis* 2(2):100–116. <https://doi.org/10.3390/prosthesis2020011>
4. Tammareddi S, Li Q (2010) Effects of material on the deployment of coronary stents. *Adv Mater Res* 123–125:315–318. <https://doi.org/10.4028/www.scientific.net/amr.123-125.315>
5. Goriainov V, Cook RM, Latham JG, Dunlop D, Oreffo ROC (2014) Bone and metal: an orthopaedic perspective on Osseointegration of Metals. *Acta Biomaterialia* 10(10):4043–4057. <https://doi.org/10.1016/j.actbio.2014.06.004>
6. Simões IG, dos Reis AC, Valente ML (2023) Influence of surface treatment by laser irradiation on bacterial adhesion on surfaces of titanium implants and their alloys: systematic review. *Saudi Dental J* 35(2):111–124. <https://doi.org/10.1016/j.sdentj.2023.01.004>
7. Kolarovszki B, Ficsor S, Frank D, Katona K, Soos B, Turzo K (2024) Unlocking the potential: laser surface modifications for titanium dental implants. *Lasers Med Sci* 39(1):162. <https://doi.org/10.1007/s10103-024-04076-1>
8. Vázquez Martínez J, Salguero Gómez J, Batista Ponce M, Botana Pedemonte F (2017) Effects of laser processing parameters on texturized layer development and surface features of Ti-6Al-4V alloy samples. *Coatings* 8(1):6. <https://doi.org/10.3390/coatings8010006>
9. Rajab FH, Al-Hamd R (2023) Surface characterization of laser processing of Ti-6Al-4V in air and Underwater. *J Russ Laser Res* 44(2):213–223. <https://doi.org/10.1007/s10946-023-10126-w>

10. McElroy S, Yang D, Reddy RG (2000) Laser processing of titanium aluminides. *J Mater Eng Perform* 9(5):506–515. <https://doi.org/10.1361/105994900770345629>
11. Baltacis K et al (2020) Physical fundamentals of biomaterials surface electrical functionalization. *Materials* 13(20):4575. <https://doi.org/10.3390/ma13204575>
12. Dekhtyar Y (2014) Weak electron emission current for characterization of nanomaterials, gas and radiation sensing towards medical applications. *Proc Est Acad Sci* 63(3):258. <https://doi.org/10.3176/proc.2014.3.09>
13. Joseph G, et al. (2003) *Scanning Electron Microscopy and X-Ray Microanalysis* (3rd ed.). Springer Science+Business, Inc.
14. JEOL, Ltd. Reflected electron, backscattered electron: Glossary: Jeol Ltd.. Reflected electron, backscattered electron | Glossary | JEOL Ltd. <https://www.jeol.com/words/semterms/20121024.072858.php#gsc.tab=0>
15. Antończak AJ, Stępak B, Koziół PE, Abramski KM (2013) The influence of process parameters on the laser-induced coloring of titanium. *Appl Phys A* 115(3):1003–1013. <https://doi.org/10.1007/s00339-013-7932-8>
16. Dekhtyar Y (2008) Photo-, dual- and Exoelectron spectroscopy to characterize nanostructures. In: *NATO Science for Peace and Security Series B: Physics and Biophysics*, pp 169–183. https://doi.org/10.1007/978-1-4020-8903-9_10
17. Shikaliev MP, Polenov BV (1981) *Open-Type Secondary Electron Multipliers and their Applications*. Energoizdat, Moscow, p 137
18. Marassi R, Nobili F (2009) Measurement methods | structural and chemical properties: scanning electron microscopy. In: *Encyclopedia of Electrochemical Power Sources*, pp 758–768. <https://doi.org/10.1016/b978-044452745-5.00071-x>



Effect of Titanium-Doped Hydroxyapatite on Biofilm Formation of *Pseudomonas aeruginosa*

Ángel Rodrigo Ortiz Juárez¹ , Jimena Mariette Robledo Dorantes¹ ,
Araceli Zapatero-Gutiérrez^{2,3} , Josué García-Ávila^{2,4} ,
José Rafael Alanis-Gómez^{2,5} , and Fabiola Hernández-Rosas^{2,3}

¹ Programa de Ingeniería Biomédica, Universidad Anáhuac Querétaro, 76246 Zibatá, Querétaro, México

² División de Ingeniería, Universidad Anáhuac Querétaro, 76246 Zibatá, Querétaro, México
fabiola.hernandezro@anahuac.mx

³ Centro de Investigación, Universidad Anáhuac Querétaro, 76246 Zibatá, Querétaro, México

⁴ Stanford University, 94305, Stanford, CA, USA

⁵ División de Investigación y Posgrado Facultad de Ingeniería, Universidad Autónoma de Querétaro, 76140 Querétaro, México

Abstract. In this study, we synthesized titanium-doped hydroxyapatite (HAP-Ti) by microwave-assisted hydrothermal method and evaluated its antibacterial activity against *Pseudomonas aeruginosa*. The synthesis was optimized by controlling parameters such as temperature, reaction time, and molar ratio of precursors to favor the homogeneous incorporation of titanium into the crystalline structure of hydroxyapatite. Structural and morphological characterization was performed by X-ray diffraction (XRD) to confirm the crystalline phase and scanning electron microscopy (SEM) to analyze the morphology and particle distribution. The antibacterial activity of HAP-Ti was evaluated by minimum inhibitory concentration (MIC) assays, cell viability assays, and biofilm inhibition in *P. aeruginosa* cultures under controlled conditions. Different concentrations of the bio-material were exposed to bacterial cultures and cell viability was determined by spectrophotometry. Biofilm inhibition was analyzed by crystal violet staining and microscopy, observing the density and distribution of adhered structures. The results showed a concentration-dependent inhibition, with a significant reduction in bacterial growth and biofilm formation in the presence of HAP-Ti. These findings highlight the potential of HAP-Ti for biomedical applications, particularly in the development of implants and prostheses with antimicrobial properties, which could contribute to the prevention of biofilm-associated infections and improve the biocompatibility of medical devices.

Keywords: Hydroxyapatite · Titanium doping · Antibacterial activity · Biofilm inhibition · Biomedical applications

1 Introduction

Infections associated with medical implants represent a major challenge in the field of biomedicine. These infections are usually caused by the formation of bacterial bio-films, which are organized microbial structures adhered to surfaces and are capable of resisting both antibiotic treatment and the immune system response [1]. Among the most prevalent pathogens in this type of infection is *Pseudomonas aeruginosa*, a gram-negative bacterium with a high capacity to form resistant biofilms on implants and medical devices [2]. To address this problem, the use of hydroxyapatite, a calcium phosphate widely used in prostheses and bone coatings due to its biocompatibility and osteoinductive properties, has been proposed [3]. In particular, titanium-doped hydroxyapatite has shown significant improvements in its biomedical performance. Its biocompatibility is optimized by presenting a composition similar to the mineral component of human bone, reducing the risk of immunological rejection. In addition, its mechanical resistance and hardness increase with the addition of titanium, while its bioactivity is favored, facilitating cell adhesion and proliferation. Likewise, its chemical stability improves, making it more resistant to degradation in the biological environment [3]. Conventionally, hydroxyapatite is used in various biomedical applications, including bone implants and prostheses, where its porous structure favors osteoconduction; coatings for medical implants, improving osteointegration; and dental applications, integrating with the maxillary and mandibular bone [4]. However, despite its advantages, it is susceptible to bacterial colonization, which can compromise the functionality of the implant. In this context, doing metals such as titanium emerges as a promising strategy to inhibit the formation of bacterial biofilms and reduce the risk of infection [5].

This research hypothesizes that titanium-doped hydroxyapatite will contribute to the reduction of *P. aeruginosa* biofilm formation, evaluating its modified properties as a possible solution to prevent bacterial colonization. This seeks to improve the performance and safety of medical implants and prostheses, offering an innovative alternative in the design of biomedical materials with greater resistance to infections.

2 Material and Methods

2.1 Synthesis of Titanium-Doped Hydroxyapatite

Titanium-doped hydroxyapatite (Ti-HAp) was synthesized by the Microwave-Assisted Hydrothermal Method, following the procedure described by Alanis-Gómez et al. (2016) [6]. As a source of titanium, 2% titanium (IV) isopropoxide (CAS Number: 546-68-9; Sigma Aldrich) was used. The reaction was carried out in a Monowave 450 reactor (Anton Paar GmbH, Graz, Austria) maintaining the same process conditions reported in the literature.

2.2 Structural and Morphological Characterization

The structural characterization of the material was performed by X-ray diffraction on an ARL EQUINOX 100 XRD diffractometer (Thermo Scientific, Waltham, MA, USA),

operated at 20 kV and 20 mA, using Cu-K α radiation ($\lambda = 0.15406$ nm) in a 2θ range of 10 to 80° with an angular step of 0.05° and a residence time of one second in each angular step. The morphological characterization was carried out by scanning electron microscopy (SEM) on a Phenom XL G2 microscope (Thermo Fisher Scientific, Waltham, MA, USA) from a sample of titanium-doped hydroxyapatite powder. Acceleration voltages of 15 kV and a magnification of x5916 were used to obtain representative micrographs of the particle morphology and distribution.

2.3 Elemental Composition Analysis

The elemental composition and purity of the titanium-doped hydroxyapatite was determined by Energy Dispersive X-ray Spectroscopy (EDS) using an Oxford AZtecOne EDS system (Oxford Instruments, Abingdon, United Kingdom). The analysis was performed under the following conditions: acceleration voltage of 20 kV, acquisition time of 60 s and a spectral resolution of 130 eV.

2.4 Preparation of Bacterial Strain

The *Pseudomona aeruginosa* strain (ATCSC 27853) was initially propagated in Nutrient Broth No. 1, a standard liquid medium (Merck Millipore, Merck Group, Germany), and subsequently isolated on blood agar (Becton Dickinson, New Jersey, USA) for growth. Bacterial suspensions and plates were incubated at 37°C for 24 hours prior to use in assays.

2.5 Minimum Inhibitory Concentration (MIC) Assay

The MIC assay was performed by the microdilution method in 96-well plates using nutrient broth No. 1 as the culture medium. A total of 100 μ L of a bacterial suspension with a density of 10^6 CFU/mL was seeded in each well (M1224-500G, HiMedia Laboratories, PA, USA). The cells were incubated at 37°C for 24 hours in the presence of different concentrations of titanium-doped hydroxyapatite (HAp-Ti) dissolved in dimethyl sulfoxide (DMSO) (Thermo Scientific™ 20688, Waltham, MA, USA), at final concentrations of 0.1, 1, 10, 100, and 200 μ g/mL. The selected concentration range was based on previously reported values for metal-doped hydroxyapatite with antimicrobial properties, ensuring coverage of both subinhibitory and inhibitory concentrations [7].

For the negative control, an untreated bacterial suspension was prepared under the same conditions but without the addition of HAp-Ti. Additionally, a vehicle control containing the maximum DMSO concentration used in the experimental wells was included to rule out any solvent-induced effects. Quantification of antimicrobial activity was performed by measuring the absorbance of microbial suspensions at 450 nm in a Multiskan FC microplate photometer (Thermo Scientific, MA, USA). All experiments were performed in triplicate, and the results were expressed as means $\pm$ standard error.

2.6 Cell Viability Assay

The cell viability assay was performed following the same procedure described for the MIC assay. However, at the end of the process, 10 μL of Alamar Blue reagent (Invitrogen™, Thermo Fisher Scientific, MA, USA) was added to each well and the plates were incubated for 4 h at 37 °C. Subsequently, the absorbance reading was performed at 570 nm in the Multiskan FC microplate reader. The experiments were performed in triplicate and the results were expressed as means $\pm$ standard deviation (SD).

2.7 Biofilm Assay

To evaluate the formation of biofilms of *P. aeruginosa*, a 96-well plate was used in which 100 μL of culture medium was placed in the first two wells, corresponding to the negative and positive controls. The negative control corresponded to culture medium inoculated with *P. aeruginosa* without treatment. For positive control, gentamicin 1 $\mu\text{g/mL}$ was used. In the third well, 1 $\mu\text{g/mL}$ of HAp-Ti in liquid culture broth was added. From the fourth well, 9 serial dilutions were made by micropipetting, starting with well 1 (1 $\mu\text{g/mL}$) and ending with well 12. Subsequently, 20 μL of a microbial suspension adjusted to 0.5 on the McFarland scale were added to each well, except the first. The plates were incubated for 24 h at 37 °C. After incubation, the biofilms were washed twice with sterile saline solution and the residual liquid was removed. They were then fixed with 200 μL of cold methanol for 5 min, the methanol was removed, and the plates were allowed to dry. The biofilms were stained with 100 μL of 1% crystal violet for 20 min and excess dye was removed by washing with 200 μL of distilled water. Once dry, the plates were observed with an inverted microscope at 20x magnification to examine the presence of biofilms. Finally, 200 μL of a 10% acetic acid solution was added and the absorbance was measured at 450 nm on the Multiskan FC photometer.

2.8 Statistical Analysis

Statistical analysis was performed using SigmaPlot software version 15.0 (SPSS Inc., Chicago, IL, USA). For comparison between groups, the ANOVA test was used. A p value < 0.05 was considered statistically significant. Graphs of means $\pm$ standard deviation were made in SigmaPlot software version 16 (SPSS Inc., Chicago, IL, USA).

3 Results

3.1 Morphological and Structural Characterization of the Biomaterial

Figure 1a shows a SEM micrograph of titanium-doped hydroxyapatite at 5916x magnification. The image shows a morphology composed of elongated crystalline structures with a cluster-like arrangement, characteristic of crystallization in the form of aggregates of acicular particles. The image reveals a rough surface with a heterogeneous distribution of the particles, suggesting a crystal growth process influenced by the presence of the doping element. Figure 1b shows the X-ray diffraction pattern of the synthesized

biomaterial, confirming the formation of the crystalline phase of hydroxyapatite. Characteristic diffraction peaks are identified at 2θ angles of approximately 25.9° , 32.9° , 39.9° and 46.7° , corresponding to the crystallographic planes (120), (211), (130) and (300), respectively. The presence of these peaks is indicative of a well-defined crystal structure, with a degree of crystallinity suitable for biomedical applications. No additional peaks attributable to significant secondary phases are observed, suggesting a homogeneous incorporation of titanium into the hydroxyapatite crystal lattice without inducing the formation of undesired compounds.

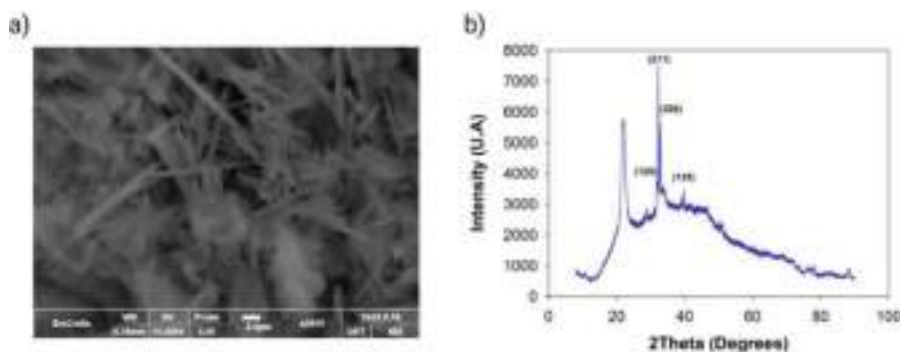


Fig. 1. Characterization of titanium-doped hydroxyapatite. (A) SEM micrograph at 5916x detailing the sample morphology (B) XRD spectrum of HAp-Ti powder, collected using Cu K α radiation, $\lambda = 0.15406$ nm.

The results of the energy dispersive X-ray (EDS) analysis of the titanium-doped hydroxyapatite sample are presented in Fig. 2. The elemental chemical mapping shown in Fig. 2a reveals a uniform distribution of the elements present in the sample. In addition, the presence of the doping titanium incorporated into the hydroxyapatite structure is evident. This mapping provides information on the spatial distribution of the elements at the microscopic level.

The corresponding SEM image, represented in Fig. 2b, shows the surface morphology of the sample on a larger scale, where the shape and size of the crystals can be observed, as well as the structural features that could have been altered by titanium doping. Figure 2c corresponds to the EDS spectrum, which shows the characteristic peaks corresponding to the elements that make up the hydroxyapatite sample, such as calcium and phosphorus, together with the peaks of titanium, confirming its incorporation into the sample.

In addition, the peaks of other elements, such as oxygen and carbon, which come from the hydroxyapatite matrix and possible surface contaminants, can be seen. These results confirm the correct incorporation of the doping element into the chemical structure of the hydroxyapatite and provide information about the distribution and morphological characteristics of the sample.

The results of the evaluation of the antibacterial activity against *Pseudomona aeruginosa* are presented in Fig. 3. Figure 3a shows graphically the results of the colorimetric analysis performed by spectrophotometry, using plate culture samples of *P. aeruginosa*

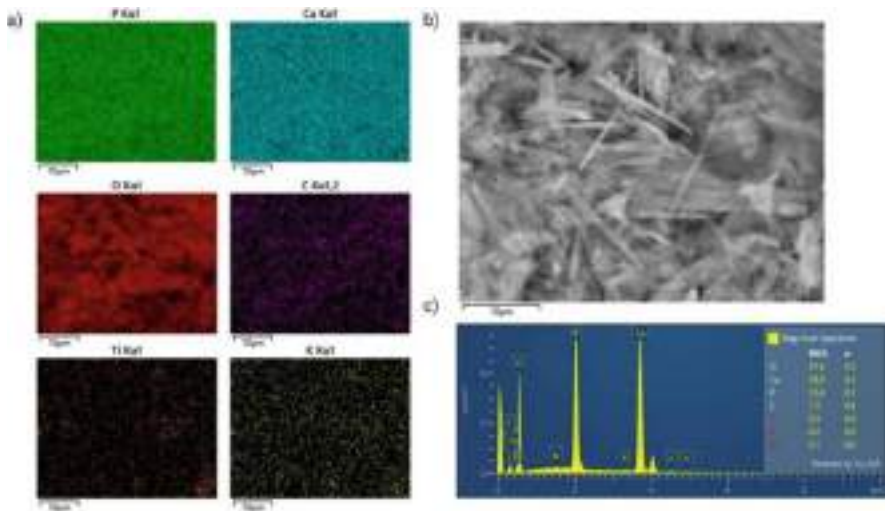


Fig. 2. Analysis of the titanium-doped hydroxyapatite sample by EDS and SEM. (a) Elemental chemical mapping. (b) SEM micrograph of the sample. (c) EDS spectrum with the table of relative abundance of the elements present in the sample.

treated with different concentrations of HAp-Ti. The intensity of the blue color of crystal violet reflects the amount of biofilm formed on the surface of the plate, which translates into an increase in absorbance. In the graph, it is observed that at higher concentrations of HAp-Ti, the absorbance values decrease, indicating a reduction in biofilm formation. This effect is more pronounced at concentrations higher than 10 $\mu\text{g/ml}$ of HAp-Ti. Figure 3b presents the results of the cell viability assay, which show the percentage reduction of the AlamarBlue reagent (resazurin) in bacteria treated with various concentrations of HAp-Ti. A greater reduction of the reagent is associated with a lower bacterial viability. The results in Fig. 3b show that at concentrations of 10 $\mu\text{g/ml}$ of HAp-Ti a significant decrease in cell viability begins to be observed, confirming the antibacterial effect of titanium-doped hydroxyapatite.

In addition, Fig. 3c presents the results of the microscopic analysis of representative samples of the biofilm assay. The biofilm of the untreated control is contrasted with the reduction observed in biofilms treated with concentrations of 100 and 200 $\mu\text{g/ml}$ of HAp-Ti, which corresponds to the reduction in absorbance observed in the graph in Fig. 3a. Taken together, these findings suggest that HAp-Ti exerts an antibacterial effect against *P. aeruginosa*, reducing biofilm formation and bacterial cell viability, which opens possibilities for its use in the development of biomaterial-based antibacterial treatments.

We can observe in Table 1 a comparative analysis of our biomaterial with the materials published in the literature. Indeed, we observed that the incorporation of titanium in the hydroxyapatite structure alters its morphology, favoring the formation of more elongated and aggregated crystalline structures, which can influence the interaction of the surface of the material with bacterial cells, as reported in the literature [10]. This structural modification seems to enhance the antibacterial properties of the material,

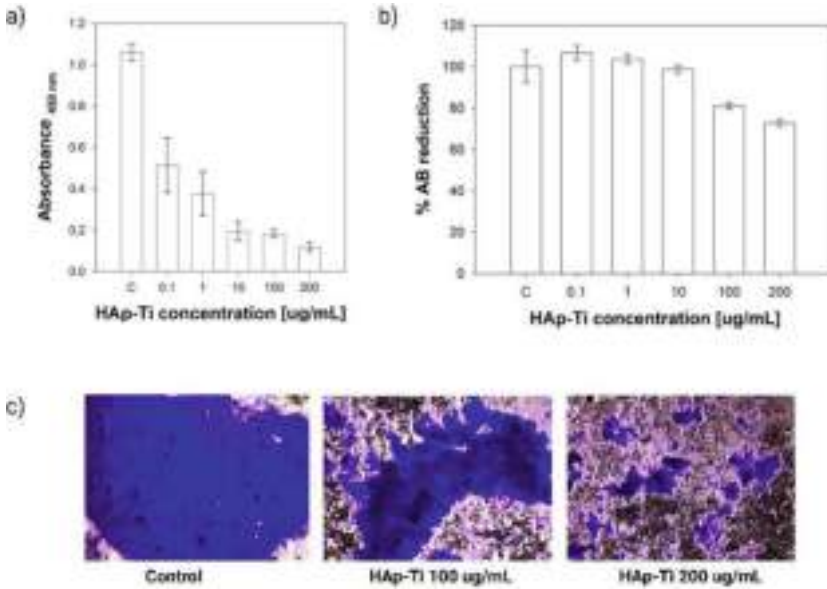


Fig. 3. Evaluation of the antibacterial activity against *Pseudomonas aeruginosa*. (A) Results of the colorimetric analysis by spectrophotometry of the biofilm assay of *P. aeruginosa* cultures treated with different concentrations of HAp-Ti. (B) Results of the cell viability assay, showing the percentage reduction of the AlamarBlue reagent in bacteria treated with various concentrations of HAp-Ti. (C) Microscopic analysis of representative samples of the plate biofilm assay observed at a magnification of 20x.

as observed in the significant reduction in biofilm formation and the decrease in cell viability of *Pseudomonas aeruginosa* at concentrations higher than 10 µg/ml of HAp-Ti. In addition, titanium doping improves the material's capacity to integrate bioactive elements and promote cell adhesion in biomedical applications, which reinforces its potential in antibacterial treatments and as a support for bone regeneration.

4 Discussion

The results obtained in this study indicate that titanium-doped hydroxyapatite (HAp-Ti) exhibits a significant inhibition in the formation of *Pseudomonas aeruginosa* bio-films, highlighting its potential application in the development of implants and medical devices with antimicrobial properties. The incorporation of titanium in the crystalline structure of hydroxyapatite seems to modify its interaction with bacteria, reducing cell adhesion and limiting proliferation under favorable conditions for biofilm formation. These findings coincide with previous studies that have reported that the inclusion of metal ions in hydroxyapatite improves its antimicrobial activity, as has been observed with silver- or copper-doped hydroxyapatite, where a reduction in the viability of different bacterial species has been demonstrated [10–12]. However, unlike these metal-based systems with high toxicity at high concentrations, titanium has excellent biocompatibility and stability, making it a safer option for long-term biomedical applications [9].

Table 1. Morphology, Structure, and Bioactivity of HAp and Titanium-Based Materials

Material	Morphology	Crystalline Structure	Antibacterial Properties	Biocompatibility	References
Pure hydroxyapatite	Spherical particles or fibers	Hexagonal, well defined	Low, without antibacterial additives	High, with good bone integration	[8]
Pure titanium	Rough, crystalline surface	Hexagonal, body centered	Moderate, with antibacterial properties	Excellent, widely used in implants	[9]
Titanium-doped hydroxyapatite	Elongated crystal structures, aggregate shape	Hexagonal	High, with significant reduction in biofilm formation	Very good, similar to pure HAp, with improvements in cell integration	[10]

Compared with other antimicrobial strategies used in implants, such as antibiotic coatings or polymers with controlled release of bactericidal agents, the advantage of HAp-Ti lies in its passive mechanism of action, where surface modification directly affects bacterial adhesion and proliferation without the need for the release of active compounds. Recent work has shown that hydroxyapatite-based materials can efficiently integrate with the bone matrix without compromising its mechanical strength, suggesting that titanium modification could offer a balance between bioactivity, structural strength and antimicrobial activity [11, 12].

In terms of practical implementation, microwave-assisted hydrothermal synthesis presents advantages in terms of purity and morphology control, allowing scalable production that would facilitate its incorporation into medical devices without the need for substantial modifications to existing manufacturing processes. However, for its eventual clinical application, it would be necessary to evaluate its behavior in *in vivo* models, considering the response of the immune system and its stability in biological fluids. The results obtained in this study reinforce the evidence on the effect of metal doping in improving the antimicrobial properties of hydroxyapatite, positioning HAp-Ti as a promising material for the prevention of implant-associated infections. However, future studies should focus on optimizing titanium concentrations and evaluating its performance in more complex models, with the aim of establishing specific parameters that guarantee its safety and efficacy in biomedical applications.

5 Conclusions

The results obtained from the experiment indicate that HAp-Ti shows an inhibitory effect on the bacterial activity of *P. aeruginosa*, with a progressive decrease in absorbance values being observed as the concentration of HAp-Ti increases. The data suggests

a positive result to validate the hypothesis that HAp-Ti can reduce the formation of biofilms on implants and prostheses. It is crucial to be able to reduce the production of biofilms on medical equipment, since it can largely prevent infections that seriously affect patients, which would substantially improve the quality of medical equipment and the quality of life of the patient. By obtaining significant values between HAp-Ti and the decrease in bacterial activity, it can be understood that this line of research is useful for the biomedical field, since it is possible to continue researching biomaterials combined with antibacterial materials that eliminate the risks of infection in implants and prostheses.

References

1. Caballero Sarmiento JS (2017) Efecto de la morfología de nanocompuestos HA/MgO/Pt en sus propiedades antibacterianas bajo activación fotocatalítica (Doctoral dissertation)
2. Solano-Gálvez SG, Paz Zarza VM, Mordani SM, Martínez-Maldonado A, Álvarez Hernández D, Vázquez-López R (2019) *Pseudomonas aeruginosa*: patogenicidad y resistencia antimicrobiana en la infección urinaria. *Re-vista Chilena de Infectología* 36:180–189. www.sochinf.cl
3. García Garduño MV, Reyes Gasga J (2006) La hidroxiapatita, su importancia en los tejidos mineralizados y su aplicación biomédica. *TIP Rev Esp Cienc Quím Biol* 90(2). www.medigraphic.com
4. Mujica KH, Castañeda-Vía JA, Landauro CV, Quispe J, Jon LYC (2021) Obtaining hydroxyapatite through biological residues for dental bone graft. *Revista Estomatológica Herediana* 31(2):111–116. <https://doi.org/10.1019/43552021000200111>
5. Paz A, Martín Y, Pazos LM, Parodi MB, Ybarra GO, González J (2011) Obtención de recubrimientos de hidroxiapatita sobre titanio mediante el método biomimético. *Revista de Metalurgia (Madrid)* 47(2):138–146. <https://doi.org/10.3989/revmetalmadrid.1009>
6. Alanís-Gómez JR et al (2016) Synthesis of micro- and nano-sized hydroxyapatite fibers through the microwave-assisted hydrothermal method. *J Nanosci Nano-technol* 16(7):7557–7566
7. Alanís-Gómez RP et al (2024) Magnesium-doped hydroxyapatite nano-fibers for medicine applications: characterization, antimicrobial activity, and cytotoxicity study. *Int J Mol Sci* 25(22):12418. <https://doi.org/10.3390/ijms252212418>
8. Dorozhkin SV (2009) Calcium orthophosphates in nature, biology and medicine. *Materials (Basel)* 2(2):399–498. <https://doi.org/10.3390/ma2020399>
9. Vallet-Regí M, Lozano D, González B, Izquierdo-Barba I (2020) Bio-materials against bone infection. *Adv Healthcare Mater* 9(13):e2000310. <https://doi.org/10.1002/adhm.202000310>
10. Ghosh R, Das S, Mallick SP, Beyene Z (2022) A review on the anti-microbial and antibiofilm activity of doped hydroxyapatite and its composites for biomedical applications. *Mater Today Commun* 31:103311. <https://doi.org/10.1016/j.mtcomm.2022.103311>
11. Besinis A, Hadi SD, Le HR, Tredwin C, Handy RD (2017) Anti-bacterial activity and biofilm inhibition by surface-modified titanium alloy medical implants following application of silver, titanium dioxide, and hydrox-yapatite nanocoatings. *Nanotoxicology* 11(3):327–338
12. O'Sullivan C, O'Neill L, O'Leary ND, O'Gara JP, Crean AM, Ryan KB (2021) Osteointegration, antimicrobial and antibiofilm activity of orthopedic titanium surfaces coated with silver and strontium-doped hydroxy-apatite using a novel blasting process. *Drug Del Transl Res* 11:702–716



Nanolayer of Monocrystalline Si Towards Nanodosimetry of Electron Radiation

Galina Boka¹, Yuri Dekhtyar², Mirko Rocca², Elizabete Skrebele²,
and Hermanis Sorokins²(✉)

¹ Oncology Center of Latvia, 4 Hipokrāta Street, Riga 1038, Latvia

² Riga Technical University, 6A Kipsalas Street, Riga 1048, Latvia

Hermanis.Sorokins@rtu.lv

Abstract. Nanodosimetry aims to measure the ionizing radiation dose absorbed in nanosized volumes. For that, photothermostimulated exoelectron emission (PTSE) of monocrystalline silicon (Si) nanolayer might be used to detect weak electrons, because their mean free path in the solid is around 100 nm. The current work aims to explore the possibility of using monocrystalline Si for nanodosimetry of high energy electron radiation. It is demonstrated that the PTSE of monocrystalline Si can be used for the identification of 6 MeV electron radiation doses 35–60 Gy. The total emitted charge of PTSE is connected with the delivered dose. The near-threshold photoelectric emission spectroscopy and Fourier-transform infrared spectroscopy (FTIR) indicate that PTSE is influenced by radiation-induced reconstructions.

Keywords: High energy electron radiation · Exoelectron emission · Monocrystalline silicon · Nanodosimetry

1 Introduction

High energy electron radiation is used for Intensity-Modulated Radiation Therapy (IMRT) [1]. Typically, electrons with an energy of 6 MeV are applied for irradiation of the tumor target that is located at a depth of several cm [2].

The dose profile has a rapid and non-linear drop between the highest and lowest levels, depending on the lateral scattering of radiation within the irradiated volume [1]. Dose deviation $\Delta D(x)$ strictly depends on a dose gradient within the dose absorbing layer [1,3]:

$$\Delta D(x) = \frac{dD}{dx} \cdot \Delta x \quad (1)$$

whereas D - dose, x - thickness.

The dose increment is important for cell survival. For instance, the dose increment of 0.05 Gy, with dose gradient 2.8 Gy/mm significantly influences the survival rate of human fibroblast (AG0-1522B), breast cancer (MDA-MB-231), prostate cancer (DU-145) and glioma (T98G) cells [4].

Considering the imperative to deliver a dose with a precision of $\pm 1\%$ [5], greater dose resolution readout is achievable by using nanolayer dosimeters, since the dose gradient is inversely proportional to the thickness of the radiation absorbing layer [3]. Under these circumstances, dose might be estimated with a detector whose thickness falls within the range of 10 to 100 nm, to detect the dose gradient at the layer having thickness ≈ 104 nm [6].

To measure the dose absorbed by the nanolayer, the latter must not be influenced by measurement probe. Therefore, the contactless dose identification is preferable.

To satisfy the need of the contactless dose measurement, the photothermostimulated exoelectron emission (PTSE) detector can be exploited [6]. Through such a technique, it is possible to detect a signal from a monocrystalline Si 100 nm thin layer that corresponds to the mean free path of the electrons (typically having a kinetic energy of 0.1 to 1.0 eV) escaped from an emitter during the detection of PTSE [6, 7].

For dose readout, the PTSE detector is simultaneously heated and exposed with photons, having an energy close to the electron work function (EWF) of the detector material. PTSE is induced due to annealing of point defects inserted by radiation [6–8].

The point defects, for instance, the vacancy - admixture complex dissociates, and the density of electrons trapped in the local states changes. As a result, the photon-induced electron quantum yield and/or electron work function alter. The detected PTSE emission current (I) of Si depends on the temperature (T) and has a maximum, indicating annealing/relaxation of point defects [6–9]. Since I depends on T, it is possible to register annealing of radiation induced imperfections. Alteration of the PTSE current on T indicates the pace of the point defect annealing, however, the total emitted charge is proportional to the amount of annealed defects.

PTSE of Si responds to irradiation with low energy protons [10], 75 keV boron [11], 50 and 100 keV argon [12], 50 keV phosphorus [13, 14] and arsenic ions [14], along with 5 MeV electrons [9, 14]. In the latter case, pace is directly connected with the concentration of defects involved in annealing reaction at current T.

Despite a wide range of radiation applied to induce PTSE of Si, the connection between delivered dose with 6 MeV electrons widely used for radiation therapy and PTSE has not been studied yet. The present article aims to explore this.

2 Materials and Methods

2.1 Specimens

The 3-inch monocrystalline Czochralski-grown Si (Cz-Si) wafers (111) doped with boron (partial resistance $10 \Omega \cdot \text{cm}$, thickness 360 μm) were used. With the diamond cutter, wafers were cut for separation into the samples, each equal

to a size $10 \times 10 \text{ mm}^2$. The equipment *S200 (Russia)* for scribing of wafers was exploited.

The approximate concentration of boron (B), oxygen (O) and carbon (C) in Cz-Si are roughly $1.4 \cdot 10^{15} \text{ atom/cm}^3$ [15], $2.5 \cdot 10^{16} \text{ atom/cm}^3$ [15], from $2 \cdot 10^{17}$ to $1 \cdot 10^{18} \text{ atom/cm}^3$ [16], respectively. However, adsorbed hydrogen (H) concentration in the Si wafer surface within $3 \mu\text{m}$ in depth might reach $10^{15} \text{ atom/cm}^3$ [17].

2.2 Irradiation of Monocrystalline Silicon

The specimens were irradiated with 6 MeV electrons using *LINAC Varian True-Beam (USA)*. Doses 12–60 Gy were delivered with a dose rate 6 Gy/min. The polymethylmethacrylate (PMMA) bolus of 13 mm thickness was used to deposit the maximum dose (D) on the surface of the specimen [2].

2.3 Measurements of Irradiated Monocrystalline Silicon

PTSE. PTSE measurement technique is described in [6, 7]. Detection of PTSE was performed in vacuum $3 \cdot 10^{-3} \text{ Pa}$. Thermostimulation was provided in a range from 20 °C to 600 °C with a heating rate 10 °C/min. For this *HeatWave Labs, Inc. Model 101303 Temperature Controller (USA)* was used.

Photostimulation was performed using an ultraviolet light source *Hamamatsu Photonics K.K. (Japan)* and diffraction monochromator *MCD-2 (Russia)*. The light wavelength according to [6, 7] was selected closely to the electron work function ($\approx 4.85 \text{ eV}$ for Si [18]).

For detection of emitted electrons, the secondary electron multiplier *VEU-6 (Russia)* was used, having background noise $\approx 0.1 \text{ electron/sec}$ [19].

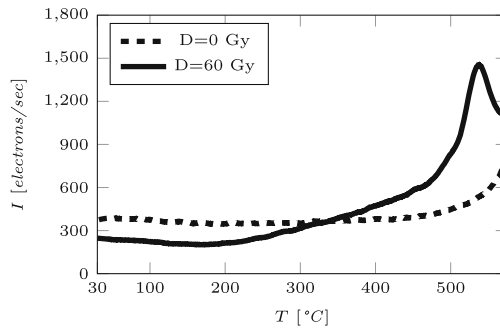


Fig. 1. PTSE of Si before ($D = 0 \text{ Gy}$) and after ($D = 60 \text{ Gy}$) irradiation

According to [6, 7] elimination of defects was indicated with the absence of the PTSE peak (Fig. 1).

Photoelectron Emission (PE). To identify reason of PTSE peak, photoelectron emission (PE) was detected on the beginning of the PTSE peak 500 °C and its maximum 550 °C. PE allowed to identify changes in the electron work function and electron emission centers density. The thermostimulation was delivered with T pace 0.3 °C/min. Typical PE spectra are demonstrated in Fig. 2.

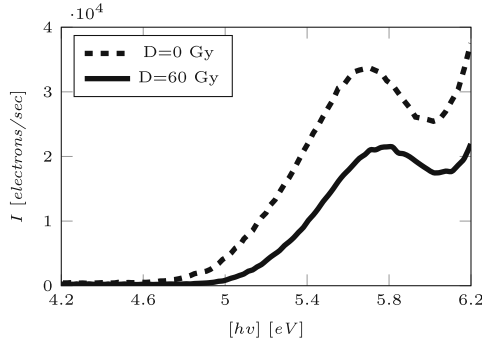


Fig. 2. PE of Si before ($D = 0$ Gy) and after ($D = 60$ Gy) irradiation at 40 °C

The photostimulation was done in a wavelength range from 300 to 200 nm (≈ 6.4 eV) with a highest possible pace 1 nm/sec. Both the change of wavelength and the heating rate were synchronized during the PE measurement according to [7]. However, before and after reaching the temperature required for the PE measurement, all PTSE conditions were used as described above.

FTIR. For FTIR spectroscopy in Attenuated Total Reflection (ATR) mode *SENSOR II FTIR Routine Spectrometer Bruker Optics (USA)* was applied within a range of 4000 to 400 cm^{-1} , having a resolution of 4 cm^{-1} . FTIR analysis was performed for the annealed samples before and after radiation treatment in atmospheric conditions.

3 Results and Discussion

3.1 PTSE of Irradiated Monocrystalline Si

It has been reported that PTSE maximum formation is observed at temperature range of 500–580 °C. In light of this, Fig. 3 demonstrates the connection of the delivered dose (D) with a total charge (Q) that has been emitted in the PTSE maximum temperature range.

The positive correlation of Q with D in accordance with [6,7] is observed for dose ≥ 36 Gy. However, a negative connection tend (the correlation coefficient is 0.97) of Q with D is observed for $D < 36$ Gy.

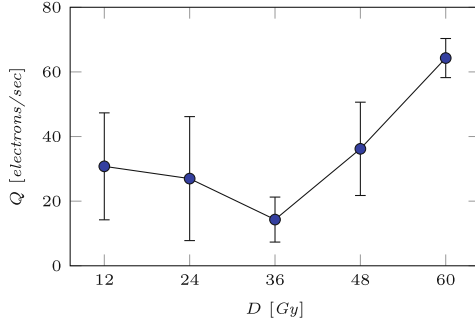


Fig. 3. Average total PTSE charge Q depending on 6 MeV electron dose D . Here and in the following sections the error bars correspond to one standard deviation.

Relevantly, following [20,21], $I(T)$ of PTSE is directly proportional to the annealing pace of the point imperfection at the corresponding T . Therefore, the non-monotonic connection of dI/dT with T characterizes that the annealing process changes (Fig. 4) [20,21].

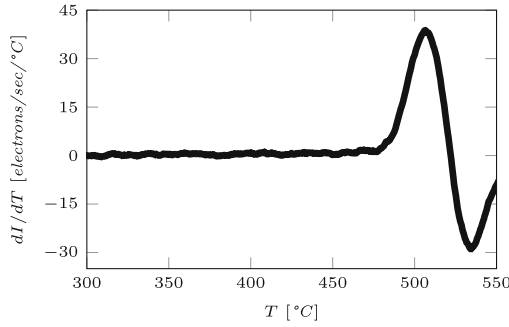


Fig. 4. dI/dT of PTSE

It has been shown that zero value of dI/dT does not depend on D . However, similar to the connection between emission Q with D , the maximum value of dI/dT (at $T \approx 500$ °C) demonstrates non-monotonic tend on D (Fig. 5): being stable at $D = 12\text{--}24$ Gy, decreasing at $D = 24\text{--}36$ Gy, and increasing at $D > 36$ Gy.

Initially, one may assume that the recorded negative correlation between Q and D could be related to the relaxation of radiation induced defects before measurement of PTSE. To verify such a hypothesis, a thin structure of the PTSE spectra, namely, dI/dT was analyzed.

However, the results of dI/dT analysis give an evidence that at $D = 12\text{--}24$ Gy the radiation induced concentration of exoemission responsible centers (single defects) is too low to influence the pace of annealing. At the next range

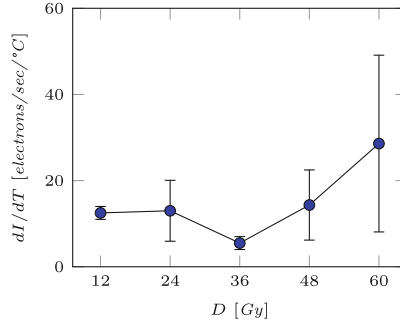


Fig. 5. dI/dT magnitude deviation on D at $T \approx 500$ °C

of D (24–36 Gy), the decrease in the annealing pace is abnormal (from the point of view of kinetic reaction theory) and could be explained by a contribution of different defects to PTSE. For doses greater than 36 Gy , the increase in defect concentration as well as the kinetics of defect annealing as a function of dose corresponds to [6,7]. Accordingly, it is possible to assume that in such a case, when $D > 36$ Gy , the defect annealing rate could also be directly proportional to the concentration of defects introduced by radiation.

3.2 PE of Irradiated Monocrystalline Si

PTSE centers were analyzed using PE spectra measurements at different T . The chosen temperatures correspond to: the baseline temperature $T_0 = 40$ °C, beginning of the PTSE peak $T_1 = 500$ °C, and peak maximum temperature $T_{max} = 550$ °C.

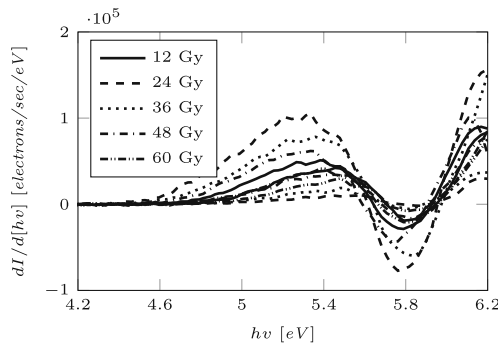


Fig. 6. $dI/d[hv]$ of PE depending on 6 MeV electron dose D at $T_1 = 500$ °C

The PE spectra detected at T_1 (as an example presented in Fig. 6) and T_{max} are similar. In both cases, the PTSE centers are identified by shifts of $dI/d[hv]$

maximums. The location of maximums of the spectra at T_1 and T_{max} depends on D (Fig. 5 and Fig. 7).

Alongside, the obtained EWF ($h\nu$ at $I = 0$) of Si at T_0 and T_1 demonstrated that in both cases the EWF changed significantly at $D = 36$ Gy (Fig. 8).

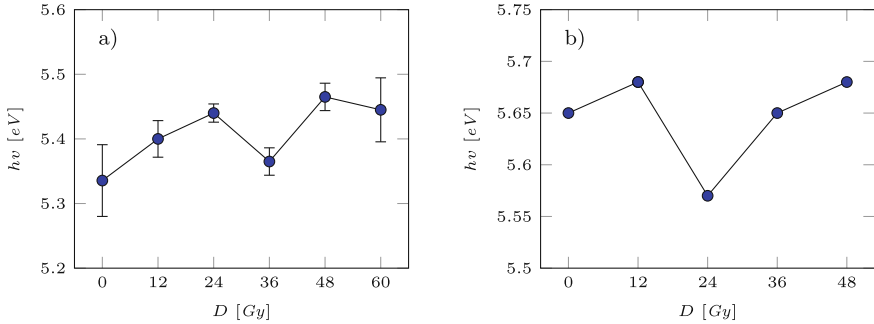


Fig. 7. Photon energy of the $dI/d[h\nu]$ maximum on D at: a) 500 °C; b) 550 °C having the error bars within the frames of the dot size.

The recorded connection of the corresponding photon energy at PE maximum energy and EWF with D is in a favor of the supposition mentioned above: increase of D changes the defect “design”. The most significant alteration of the defects appears at $D = 36$ Gy, as the $h\nu(D)$ has a sharp bend.

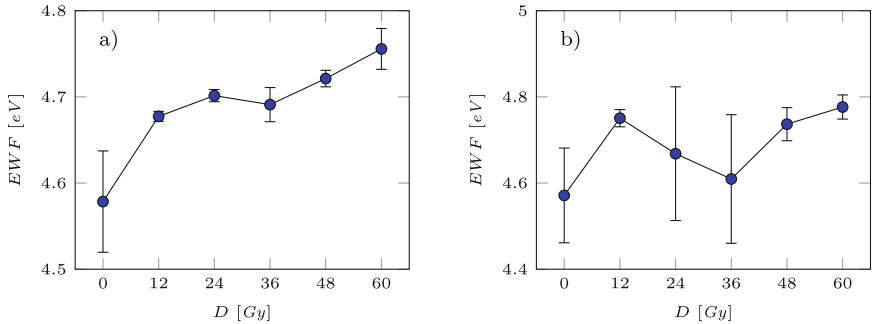


Fig. 8. EWF of Si on D at a) 40 °C and b) 500 °C

3.3 FTIR of Irradiated Monocrystalline Si

The FTIR spectra (the baseline was extracted) indicated the presence of Si-H, Si-O, and Si-OH couples observed at 2390 cm^{-1} , 1105 cm^{-1} , and 890 cm^{-1} ,

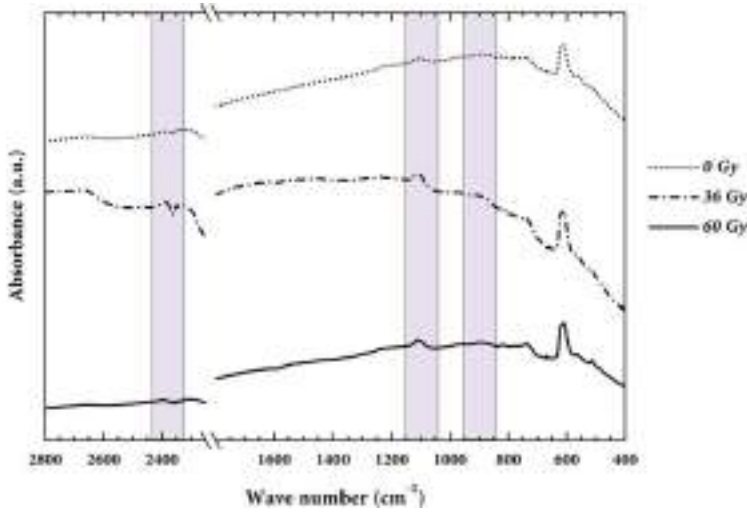


Fig. 9. FTIR of Si before and after irradiation

respectively [23,24]. As an example FTIR spectra of Si depending on doses 0 *Gy*, 36 *Gy* and 60 *Gy* is represented in Fig. 9.

The heights (*H*) of these peaks demonstrated significant non-monotonic connection with *D*. In particular, at *D* = 36 *Gy* the bent of the line *H*(*D*) is observed (Fig. 10).

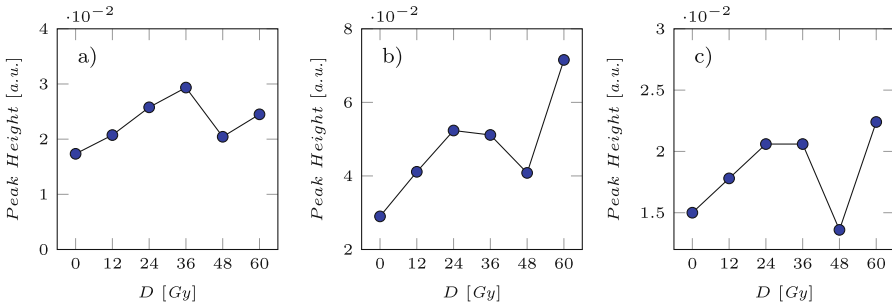


Fig. 10. *H*(*D*) of a) Si-H, b) Si-OH and c) Si-O

Each of the FTIR identified bonds involves Si-. Moreover, the results also indicate that the radiation-produced defects in the monocrystalline Si depend on *D*.

Notably, when the dependence trend of determined Si exoemission parameters - PTSE charge *Q*, magnitude *dI/dT*, emission centers, and materials' EWF at PTSE peak temperatures - on dose is compared with one related to changes in

chemical composition, a similar behavior has been observed. In other words, analyzed PTSE characteristics exhibit changes in the monotonicity of their dependence on the dose at a threshold value of 36 *Gy*. So, one could assume that the radiation-induced structural reconstructions in the semiconductor may also play a crucial role in the observed PTSE effect.

4 Conclusions

P-type monocrystalline Si irradiated with 6 MeV electrons demonstrates that:

1. Nanolayer of monocrystalline Si could be used for nanodosimetry of electron radiation.
2. A surface nanolayer of monocrystalline Si has the potential to be used for 6 MeV electron radiation nanodosimetry, identifying PTSE charge Q for doses from 36 to 60 *Gy* with measurement precision of ± 9 *electrons/sec*. Alternatively, the dose readout could be done by estimating photon energy corresponding to $dI/d[h\nu]$ maximum of PE spectra (when $T = 500$ °C) from 0 to 24 *Gy* and 36 to 60 *Gy* with measurement precision of ± 0.033 eV and ± 0.031 eV, respectively. Lastly, for dose estimation in the range of 0 to 60 *Gy* one may determine the EWF of Si at $T = 40$ °C, precision ± 0.02 eV.
3. Photothermostimulated exoelectron emission is suitable for the readout of the delivered doses by 6 MeV electron radiation.

Acknowledgements. The authors are much obliged for the cooperation with A/S “RD ALFA Mikroelektronikas Departaments” and the Oncology Center of Latvia in the matters regarding sample preparation and irradiation, respectively.

The research was supported with the European Commission project 101008571 — PRISMAP — H2020-INFRAIA-2018-2020/H2020-INFRAIA-2020-1.

References

1. Li JS, Lin T, Chen L, Price RA Jr, Ma CM (2010) Uncertainties in IMRT dosimetry. *Med. Phys.* 37(6):2491–2500. <https://doi.org/10.1118/1.3413997>
2. Elledge, C.R., Alcorn, S.R.: Treatment planning in palliative radiotherapy. In: *Palliative Radiation Oncology*, pp. 65–74. Elsevier (2023). <https://doi.org/10.1016/b978-0-323-87688-9.00017-9>
3. Ramachandran P et al (2009) Impact of different CT slice thickness on clinical target volume for 3D conformal radiation therapy. *Med. Dosim.* 34(1):36–41. <https://doi.org/10.1016/j.meddos.2007.09.002>
4. Trainor C et al (2012) Cell survival responses after exposure to modulated radiation fields. *Radiat. Res.* 177(1):44–51. <https://doi.org/10.1667/RR2656.1>
5. International Atomic Energy Agency: *Radiation Oncology Physics: A Handbook for Teachers and Students*. Technical Report Series No. 896, IAEA, Vienna (2007). https://www-pub.iaea.org/MTCD/Publications/PDF/te_896_prn.pdf
6. Dekhtyar Yu (2013) Emission of weak electrons: dosimetry of nanovolumes. *Radiat. Meas.* 55:34–37. <https://doi.org/10.1016/j.radmeas.2013.01.009>

7. Dekhtyar Yu (2014) Weak electron emission current for characterization of nano-materials, gas and radiation sensing towards medical applications. *Proc. Estonian Acad. Sci.* 63(3):346–354. <https://doi.org/10.3176/proc.2014.3.09>
8. Dekhtyar, Y.: Exoelectron spectroscopy of silicon defects. *Surface: Phys. Chem. Mech.* **9**, 57–67 (1993)
9. Dekhtyar, Y.D., Sagalovich, G.L.: Photothermostimulated Exoelectron emission from crystalline silicon containing edge dislocations and point defects. *Physica status solidi (a) - Appl. Mater. Sci.* **102**(2), 665–677 (1987). <https://doi.org/10.1002/pssa.2211020227>
10. Zdanovich, A.G., Kryuk, V.I., Mints, R.I., Milman, I.I.: Exo-electron emission of silicon and germanium after bombardment by protons of low energy. Görlich (ed), Berlin, Boston: De Gruyter **8**(1), 363–366 (1971). <https://doi.org/10.1515/9783112472705-044>
11. Alekseev, A.N., Bystrov, V.S., Dekhtyar, Y.D., Litvinov, A.P., Sagalovich, G.L., Shendrik, A.V.: Influence of VV-centers in silicon on its photothermally stimulated emission. *Semiconductor Phys.* **1**, 169–170 (1984)
12. Dekhtyar YD, Noskov VA, Savvaitova YA, Sagalovich GL (1990) Formation and annealing of vacancy complexes in the surface layer of single-crystalline silicon irradiated with medium-mass ions. *Phys. Technol. Semiconductors* 24(3):492–496
13. Dekhtyar YD, Sagalovich GL (1990) Kinetics of defect formation in silicon upon implantation of phosphorus ions into it. *Phys. Technol. Semiconductors* 24(4):757–758
14. Dekhtyar, Y.: Exoelectron emission testing of technology inserting point defects into semiconductors. *SPIE Proc.* **2334**, 120–126 (1994). <https://doi.org/10.1117/12.186740>
15. National Institute of Standards and Technology (NIST): Radiation Measurements in Calibration Laboratories. NBS Special Publication 400-64 (1985). <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication400-64.pdf>
16. Norwegian University of Science and Technology (NTNU): Radiation Effects in Silicon-Based Materials. Fulltext Thesis. NTNU Open Repository (2017). https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2494322/14395_FULLTEXT.pdf?sequence=1&isAllowed=y
17. Tokuda Y, Kato I, Ohshima H, Hattori T (1994) Observation of hydrogen in commercial Czochralski-grown silicon wafers. *Semiconductor Sci. Technol.* 9:1733–1735. <https://doi.org/10.1088/0268-1242/9/9/026>
18. Novikov A (2010) Experimental measurement of work function in doped silicon surfaces. *Solid-State Electron.* 54:8–13. <https://doi.org/10.1016/j.sse.2009.09.005>
19. Lampes et Tubes: Photomultipliers - Detailed Overview. Accessed <https://lampes-et-tubes.info/pm/pm179.php?l=>
20. Dekhtyar YD, Vinyarskaya YA (1994) Exoelectron analysis of amorphous silicon. *J. Appl. Phys.* 75:4201–4207. <https://doi.org/10.1063/1.356005>
21. Breslav IB, Dekhtyar YD, Sagalovich GL (1993) Surface physics. *Chem. Mech.* 9:68 (in Russian)
22. LibreTexts: Oxidation of Silicon. In: Barron, A.R. (ed.), *Chemistry of the Main Group Elements*. [https://chem.libretexts.org/Bookshelves/Inorganic.Chemistry/Chemistry_of_the_Main_Group_Elements_\(Barron\)/07%3AGroup14/7.11%3A_Oxidation_of_Silicon](https://chem.libretexts.org/Bookshelves/Inorganic.Chemistry/Chemistry_of_the_Main_Group_Elements_(Barron)/07%3AGroup14/7.11%3A_Oxidation_of_Silicon)

23. Jeurink, J., Seiffe, J., Hofmann, M., Rentsch, J., Preu, R.: Plasma deposition of amorphous silicon dopant sources. In: Proceeding of 26th EU PVSEC 2011, pp. 5–9. <https://doi.org/10.4229/26thEUPVSEC2011-2BV.1.37>
24. Balaji N, Park C, Chung S, Ju M, Raja J, Yi J (2016) Effects of low temperature anneal on the interface properties of thermal silicon oxide for silicon surface passivation. *J. Nanosci. Nanotechnol.* 16(5):4783–4787. <https://doi.org/10.1166/jnn.2016.12178>



Tunable Isotropic TPMS-Based 3D-Architected Scaffolds for Diabetic Bone Regeneration

Fátima Morales-Gómez¹ , Fabiola Hernández-Rosas^{1,2} ,
Araceli Zapatero-Gutiérrez^{2,3} , Erick Ramírez-Cedillo⁴ ,
and Josué García-Ávila^{3,5}

¹ Ingeniería Biomédica, División de Ingenierías, Universidad Anáhuac Querétaro, Santiago de Querétaro 76246, México

² Centro de Investigación, Universidad Anáhuac Querétaro, Santiago de Querétaro 76246, México

³ División de Ingenierías, Ingeniería Mecánica para la Innovación, Universidad Anáhuac Querétaro, Santiago de Querétaro 76246, México
josue.garciaavila@anahuac.mx

⁴ School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey 64700, Nuevo León, México

⁵ Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA

Abstract. Diabetes mellitus significantly compromises bone health, increasing the risk of fractures, osteoporosis, and limb amputations due to metabolic dysfunctions. Effective bone regeneration is critical for managing complex fractures and bone-related complications, with over two million bone grafts performed annually. This research focuses on designing and developing personalized 3D-architected scaffolds to enhance bone regeneration in patients with diabetes using computational simulation techniques. These scaffolds are tailored to replicate the mechanical and structural properties of natural bone by optimizing key features such as anisotropy, porosity, and nutrient mass flow. The study highlights the transformative potential of personalized 3D-architected scaffolds in addressing critical challenges associated with diabetic bone health. Customizing scaffold designs to meet the specific needs of patients with diabetes aims to reduce postoperative complications, improve healing outcomes, and enhance overall patient quality of life. Future work will refine scaffold designs and validate their performance in diabetes-specific clinical *in vitro* and *in vivo* models.

Keywords: Scaffold · Bone Regeneration · Diabetes Mellitus

1 Introduction

Diabetes mellitus significantly compromises bone health, increasing the risk of fractures, osteoporosis, and limb amputations [1]. Effective bone regeneration is critical for managing complex fractures and bone-related adverse complications, with over two million bone grafts performed annually [2]. In people with diabetes, metabolic dysfunctions

impair bone healing and reduce regenerative capacity, complicating recovery. Hydroxypapatite, a natural bioceramic found in bone and teeth, is widely used in implants due to its biocompatibility, osteoconductivity, and ability to integrate with bone tissue [3]. The interplay between insulin regulation and bone structure is crucial in understanding diabetes-induced bone pathologies. Diabetes alters bone metabolism and microarchitecture, diminishing bone quality and increasing fracture susceptibility [4] (Fig. 1A). Type 1 diabetes reduces bone mass and strength, elevating fracture risk fivefold [5], while Type 2 diabetes, despite normal bone mass, leads to higher fracture rates due to compromised bone quality and increased fall incidents from complications like retinopathy and neuropathy [6].

The potential loss of a lower limb is a common condition in patients with diabetes, who have a 30 times higher lifetime risk of undergoing amputation compared to non-diabetic individuals [7]. This increases the initial recovery time and affects initial mobility (Fig. 1B). Among the surgical procedures used to improve the quality of life of individuals requiring amputation is the Ertl procedure, also known as the transtibial osteomyoplastic amputation technique. This technique focuses on improving traditional amputation methods, which often created non-physiological conditions where the residual limb lacked the structure, function, and biomechanics of a normal limb, leading to uncoordinated movements, loss of support, circulation problems, and passive functionality of the stump [8]. The Ertl technique aims to create a functional stump by addressing these limitations through the formation of a tibiofibular bridge (Fig. 1C), which helps stabilize the distal tibia and fibula (Fig. 1D) while optimizing soft tissues and circulation. This allows for better body weight-bearing transmission on the stump, making it more functional and less painful.

However, this technique is not ideal for patients with diabetes. A portion of the tibiofibular bridge is typically used to create the tibial bone or periosteum [8]. Considering the properties of diabetic bone, this could have the opposite effect of what the bridge aims to achieve; it might lead to fissures or fractures due to the forces exerted on it or delayed union of the bone bridge caused by osteoporosis in patients with diabetes. Additionally, wound healing problems may arise due to reduced blood perfusion, as well as intraoperative and postoperative complications if blood glucose levels are not well controlled [9]. 3D-architected scaffolds offer a novel approach to mimic the natural bone environment and promote regeneration. In the past, authors have shown that the use of light-based additive manufacturing enables the precise customization of scaffolds for dynamic loading conditions [10, 11].

This project aims to design and develop personalized 3D-architected scaffolds optimized for mass flow (permeability) and structural integrity (tunable anisotropy). By customizing the porous scaffold architecture according to each patient's specific bone characteristics, future efforts can be enabled to enhance osteoconductivity and facilitate seamless integration with the existing bone matrix. The focus is on its application within the Ertl surgical technique, where the scaffold would replace the tibiofibular bridge, with expected outcomes such as reduced postoperative complications, improved healing times, enhanced quality of life for patients with diabetes, and a potential improvement in the bone structure of the bones adjacent to the scaffold. This research addresses a

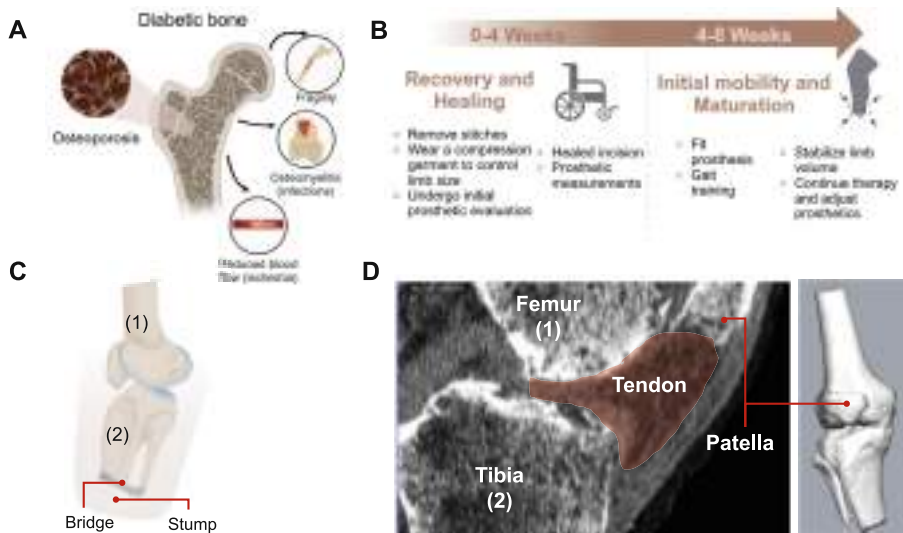


Fig. 1. (A) Diabetic bone with osteoporosis and its consequences. (B) Post-amputation timeline illustrating recovery (0–4 weeks) and initial mobility (4–8 weeks) phases. (C) Schematic of the Ertl bridging approach, depicting the tibio-fibular stump ‘bridge’ and its stabilization. (D) CT cross-section of the knee joint highlighting key anatomical landmarks—femur (1), tibia (2), patella, and the tendon region.

critical clinical need, particularly in regions with a high prevalence of diabetes, thereby contributing to advancements in bone tissue engineering and regenerative medicine.

2 Materials and Methods

The homogenization of materials was performed using nTopology® (nTopology, New York City, NY, USA). Three scaffold structures with different surface-based topologies were created, varying porosity and surface area—essential characteristics for bone simulation—as well as permeability and anisotropy. Three different Triply Periodic Minimal Surfaces (TPMS) were used to design the scaffold. These are three-dimensional surfaces with zero mean curvature that repeat periodically in all spatial directions. They are often described by implicit equations of the form $f(x, y, z) = \beta$. Where $\beta = 0$ recovers the standard mid-surface. This offset shifts the iso-surface, effectively “inflating” or “shrinking” the TPMS relative to its central geometry. For example:

$$\text{Gyroid} : \sin(x)\cos(y) + \sin(y)\cos(z) + \sin(z)\cos(x) = \beta \tag{1}$$

For the computational simulations, 20 mid-surface offset data points were used for SplitP, Gyroid, and Diamond structures, with values varying depending on the piece, ranging from a negative value to its corresponding positive extreme (e.g., -1, ..., 1) (Fig. 2A). Permeability and porosity are calculated through a macroscopic approach

using COMSOL Multiphysics® 6 (COMSOL, Burlington, MA, USA), the permeability κ is derived using Darcy's law, with the relationship:

$$\kappa = \frac{u_{out} \times \mu_f}{\Delta p / L} \quad (2)$$

where u_{out} is the outlet velocity, μ_f is the dynamic viscosity of the fluid, Δp is the pressure difference across the model, and L is the characteristic length of the porous structure. The anisotropy index A_z is the Zener anisotropy index and is given by:

$$A_z = \frac{2C_{44}}{C_{11} - C_{12}} \quad (3)$$

Here, C_{11} , C_{12} , and C_{44} are components of the elastic stiffness tensor. The Universal Anisotropy Index A_u is a dimensionless measure that quantifies the degree of anisotropy in a material. It is defined as:

$$A_U = 5 \frac{G_V}{G_R} + \frac{K_V}{K_R} - 6 \quad (4)$$

where G and K are the shear and bulk moduli in the Voigt scheme (upper bound) and the Reuss schema (lower bound).

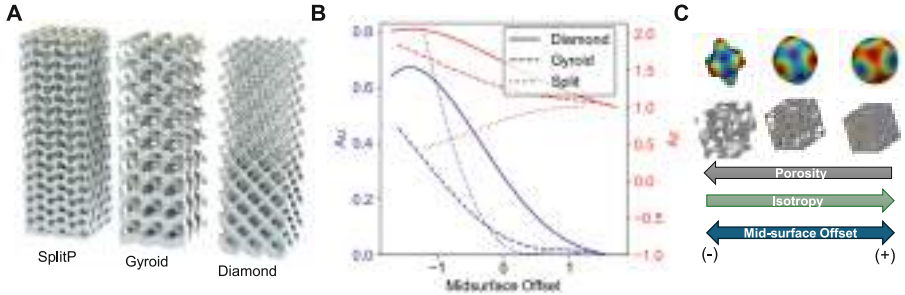


Fig. 2. (A) Examples of 3D-architected lattice scaffolds (SplitP, Gyroid, and Diamond) designed for tunable functionally graded mechanical and biological performance. (B) Relationship between the acoustic-impedance ratio r (A_U , left axis) and the morphological factor A_Z (right axis) versus mid-surface offset for Diamond, Gyroid, and Split lattice architectures. (C) 3D representations showing how porosity, isotropy, and mid-surface offset (from negative to positive) affect each lattice design. Spherical color maps illustrate local anisotropy variations for different offsets.

2.1 Patient Case Study

As reference we use a patient profile who is a 63-year-old individual, with the amputation resulting from dysvasculature due to type 2 diabetes. The amputation occurred on the right side, and the stump's shape is cylindrical. The external geometry of the stump was obtained following the method outlined by Garcia et al. [10]. This technique for capturing the stump's geometry is integral for simulating the fit and function of transtibial implants and optimizing their design, particularly when using 3D-printed, tunable isotropic porous metamaterials for orthopedic applications.

3 Results and Discussion

We studied the relationship between the mid-surface offset (β) and the anisotropy indices A_U and A_Z for three lattice structures (Fig. 2B): Diamond (solid line), Gyroid (dashed line), and Split (dotted line). Both Gyroid and Split reach near-isotropic values ($A_U = 0$) more rapidly than Diamond. Notably, the Split structure achieves lower A_Z values (ranging from 0.5 to 1), whereas Diamond and Gyroid typically progress from higher values (around 2 approaching 1). The mid-surface offset is a critical parameter in controlling geometric distortion and, therefore, the anisotropic behavior of these lattices. When $\beta > 0$, the iso-surface expands on one side, reducing the local cross-sectional area of the pores; conversely, for $\beta < 0$, it contracts toward the opposite side, altering pore shape and size in the inverse manner. Even small changes in β can substantially modify pore morphology, as seen in an offset Gyroid, where the “neck” regions between cells become narrower or wider, affecting both stress transmission and stiffness in different directions. In addition, varying β affects the porosity of the lattice: larger positive offsets lead to a closer structure (Fig. 2C).

From a topological perspective, these changes can be characterized via Minkowski functionals (e.g., mean curvature, Gaussian curvature) since shifting the surface locally modifies the 3D connectivity of pores. Although relatively unexplored in the literature, this concept offers the potential for “programmable” anisotropy: by imposing non-uniform offsets $\beta(x)$ within a single lattice, one can introduce gradients in thickness or porosity to create novel porous structures [12]. Such functionally graded designs enable transitions from stiffer regions to more compliant ones within the same component, presenting novel opportunities for tailored mechanical properties. We also analyzed the Relative Surface Area Index (RSAI), apparent density (ρ_{app}) and the anisotropy parameter (A_U) for three TPMS-based lattices according to the variation of the mid-surface offsets (Fig. 3A). Larger RSAIs indicate a greater surface area. In many cases, this can favor growth and proliferation when the structures are used as scaffolds for bone regeneration [13]. Concurrently, A_U decreases closing to near-isotropic values ($A_U = 0$) with rising RSAI (Fig. 3B), especially for the Gyroid and SplitP lattices, suggesting that higher surface regions can achieve more isotropic mechanical properties.

The permeability and anisotropy values measured for the SplitP, Gyroid, and Diamond shapes at various mid-surface offsets exhibited distinct flow patterns. Notably, the Diamond design demonstrated lower permeability (on the order of $1 \times 10^{-8} \text{ m}^2$) at negative offsets, indicating a closed-pore architecture that effectively reduces fluid flow as the mid-surface value is tuned. Similarly, both SplitP and Gyroid structures showed reduced permeability at positive offsets, which can be attributed to narrower channels and decreased effective porosity. The Gyroid structure, in particular, tends to distribute open pores within more defined channels. SplitP has an intricate double-surface architecture that facilitates a zigzag flow pattern, enhancing the distribution of fluid through the scaffold. This unique flow path not only improves nutrient transport but also supports uniform cell proliferation across the scaffold. Conversely, the Diamond structure at elevated mid-surface values features smaller pores, as evidenced by the flow patterns illustrated in Fig. 4A.

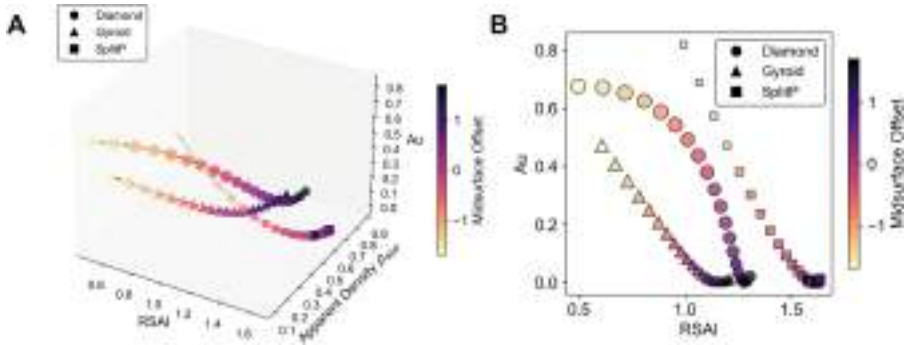


Fig. 3. (A) 3D plot of apparent density (ρ_{app}), the Relative Surface Area Index (RSAI), and the anisotropy parameter (A_U) for the three TPMS-based lattice designs. The marker size reflects the morphological factor (A_Z). RSAI is a normalized measure of total surface area relative to a reference cubic surface area of $6w^2$, where w is the unit cell width. (B) 2D projection of A_U versus RSAI for the same lattices, with color again representing the mid-surface offset.

Regarding anisotropy, the Diamond shape reached values exceeding 2.0 at negative offsets, consistent with previously reported high anisotropy indices in trabecular-like structures [14]. Utilizing Kernel Density Estimation (KDE), regions of high and low point concentration were identified within the permeability–anisotropy plane. This analysis facilitated the detection of clusters, subpopulations, and outliers among the different geometric configurations (Fig. 4B). Such a comprehensive visualization of the data space not only aids in validating or refuting existing hypotheses but also provides deeper insights into the distribution and segregation of various geometric architectures.

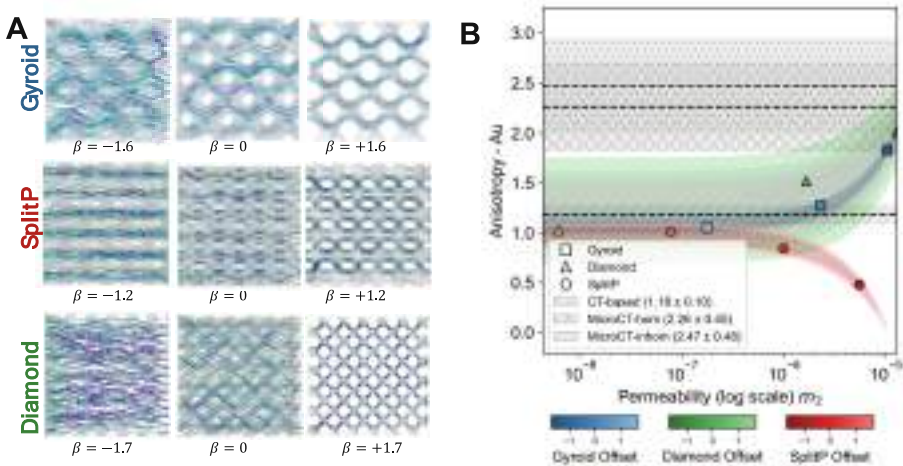


Fig. 4. (A) Representative 3D morphologies for Gyroid, SplitP, and Diamond surfaces at selected offset (β) values. (B) Permeability–Anisotropy (log scale) plot with 2D KDE shading for each structure, colored by offset.

These findings align with recent studies employing triply periodic minimal surfaces (TPMS) to optimize bone regeneration [15, 16]. Such studies demonstrate that precise variations in the TPMS architecture—particularly in the mid-surface offset—enable the modulation of critical mechanical and functional parameters, including anisotropy, stiffness, and permeability. This parametric control is crucial for ensuring effective nutrient delivery and adequate load transfer, both of which are pivotal in complex clinical settings. Comparison of our results with the literature confirms the importance of selectively adjusting density and pore connectivity to promote cell proliferation and achieve stable tissue integration. This strategy is especially pertinent for people with diabetes, who commonly present higher nonunion rates and fragility fractures, as well as a less favorable biological environment for healing [17]. By designing scaffolds with an appropriate anisotropic profile and density gradient, more robust osseointegration can be fostered, thereby mitigating complications associated with chronic hyperglycemia, such as diminished bone quality and prolonged healing times [18].

4 Conclusion

Based on the results obtained, we can conclude that customizing Triply Periodic Minimal Surfaces (TPMS) by adjusting mid-surface offset values effectively modifies both permeability and anisotropy in scaffold structures. This tunability is crucial for applications in orthopedics and tissue engineering, enabling the development of personalized biocompatible implants with optimized mechanical properties. Specifically, in the context of the Ertl transtibial osteomyoplastic amputation technique, tailored TPMS scaffolds can provide the necessary strength and permeability to promote proper osseointegration. By customizing the porous architecture to each patient's unique bone characteristics, this approach has the potential to improve healing times and reduce postoperative complications. Additionally, the distinct responses of Diamond, Gyroid, and SplitP lattice structures to mid-surface modifications highlight the versatility of TPMS in achieving desired mechanical performance and supporting uniform cell proliferation in future studies.

However, despite these promising findings, there are limitations that must be addressed in future research. It is essential to evaluate the biological response of the customized scaffolds through comprehensive *in vitro* and *in vivo* studies to confirm their biocompatibility and effectiveness in promoting bone regeneration. Long-term mechanical testing under physiological loads is also necessary to ensure the durability and reliability of the scaffolds in clinical settings. Future directions should include the integration of ceramic-based biomaterials (e.g., hydroxyapatite) to enhance the biomechanical and biocompatible properties of the scaffolds, as well as the development of advanced high-resolution 3D additive manufacturing technologies to overcome current challenges in tissue engineering. Additionally, exploring the scaffold's role not only as a structural component in the Ertl technique but also as a preventive tool against osteoporosis in adjacent bones could further improve patient outcomes. By addressing these aspects, custom 3D-architected scaffolds hold great potential for reducing complications and enhancing the quality of life for patients with diabetes, representing a significant advancement in bone tissue engineering and regenerative medicine.

References

1. Jackuliak P, Payer J (2014) Osteoporosis, fractures, and diabetes. *Int J Endocrinol* 2014:820615. <https://doi.org/10.1155/2014/820615>
2. Archunan MW, Petronis S (2021) Bone grafts in trauma and orthopaedics. *Cureus* 13:e17705. <https://doi.org/10.7759/CUREUS.17705>
3. Mondal S et al (2023) Hydroxyapatite: a journey from biomaterials to advanced functional materials. *Adv Colloid Interface Sci* 321:103013. <https://doi.org/10.1016/J.CIS.2023.103013>
4. De Waard EAC, Van Geel TACM, Savelberg HHCM, Koster A, Geusens PPMM, Van Den Bergh JPW (2014) Increased fracture risk in patients with type 2 diabetes mellitus: an overview of the underlying mechanisms and the usefulness of imaging modalities and fracture risk assessment tools. *Maturitas* 79:265–274. <https://doi.org/10.1016/J.MATURITAS.2014.08.003>
5. Hofbauer LC et al (2022) Bone fragility in diabetes: novel concepts and clinical implications. *Lancet Diab Endocrinol* 10:207–220. [https://doi.org/10.1016/S2213-8587\(21\)00347-8](https://doi.org/10.1016/S2213-8587(21)00347-8)
6. Poiana C, Capatina C (2017) Fracture risk assessment in patients with diabetes mellitus. *J Clin Densitom* 20:432–443. <https://doi.org/10.1016/J.JOCD.2017.06.011>
7. Sexton AT, Fleming LL (2022) Lower extremity amputation. In: *Medical Management of the Surgical Patient: A Textbook of Perioperative Medicine*, pp 741–743. <https://doi.org/10.1017/CBO9780511544590.119>
8. Osteomyoplastic Transtibial Amputation: The Ertl Technique (2016). <https://doi.org/10.5435/JAAOS-D-15-00026>
9. Taylor BC, Poka A (2011) Osteomyoplastic transtibial amputation: technique and tips. *J Orthop Surg Res* 6:1–4. <https://doi.org/10.1186/1749-799X-6-13>
10. García-Ávila J et al (2024) Dynamic topology optimization of 3D-Printed transtibial orthopedic implant using tunable isotropic porous metamaterials. *J Mech Behav Biomed Mater* 153:106479. <https://doi.org/10.1016/J.JMBBM.2024.106479>
11. Peto M, García-Ávila J, Rodríguez CA, Siller HR, da Silva JVL, Ramírez-Cedillo E (2024) Review on structural optimization techniques for additively manufactured implantable medical devices. *Front Mech Eng* 10:1353108. <https://doi.org/10.3389/FMECH.2024.1353108/BIBTEX>
12. García-Ávila J, Cuan-Urquizo E, Ramírez-Cedillo E, Rodríguez CA, Vargas-Martínez A (2022) Novel porous structures with non-cubic symmetry: synthesis, elastic anisotropy, and fatigue life behavior. <https://doi.org/10.1177/10812865221104236>
13. Cheng A et al (2019) Advances in porous scaffold design for bone and cartilage tissue engineering and regeneration. *Tissue Eng Part B Rev* 25:14. <https://doi.org/10.1089/TEN.TEB.2018.0119>
14. Rana M, Karmakar SK, Verdonshot N, Roychowdhury A (2024) Prediction of micro-scale bone adaptation of human trabecular bone under different implanted conditions. *J Mech Behav Biomed Mater* 160:106747. <https://doi.org/10.1016/J.JMBBM.2024.106747>
15. Agarwal R, Malhotra S, Gupta V, Jain V (2023) Three-dimensional printing of triply periodic minimal surface structured scaffolds for load-bearing bone defects. *Polym Eng Sci* 63:972–985. <https://doi.org/10.1002/PEN.26258>
16. Liu Q et al (2024) Functionally graded triply periodic minimal surface scaffold of HA-Al₂O₃ via Vat photopolymerization 3D printing. *Addit Manuf* 84:104117. <https://doi.org/10.1016/J.ADDMA.2024.104117>
17. Stewart SK (2019) Fracture non-union: a review of clinical challenges and future research needs. *Malays Orthop J* 13:1. <https://doi.org/10.5704/MOJ.1907.001>
18. Murray CE, Coleman CM (2019) Impact of diabetes mellitus on bone health. *Int J Mol Sci* 20:4873. <https://doi.org/10.3390/ijms20194873>



Biomedical Potential of the Methanolic Leaf Extract of *Verbesina persicifolia*: Antimicrobial Properties and Applications in Wound Healing

Carlos Alberto López-Rosas¹ , Seojin Han² , Margarita V. Saavedra Vélez¹ ,
José Rafael Alanís Gómez² , and Fabiola Hernández-Rosas^{2,3}

¹ Facultad de Química Farmacéutica Biológica, Universidad Veracruzana, 91090
Xalapa-Enríquez, Veracruz, México

² Escuela de Ingeniería Biomédica, Universidad Anáhuac Querétaro, 76246 Zibatá, Querétaro,
México

fabiola.hernandezro@anahuac.mx

³ Centro de Investigación, Universidad Anáhuac Querétaro, 76246 Zibatá, Querétaro, México

Abstract. Extracts from the leaves of plants of the genus *Verbesina* have shown promising biomedical potential, especially for their antimicrobial and anti-inflammatory properties and their application in wound healing. In this study, we analyzed the biological properties of the methanolic extract of *Verbesina persicifolia* using the minimum inhibitory concentration (MIC) assay to determine its antimicrobial activity against various pathogenic bacteria. In addition, cell viability was evaluated using the Alamar Blue assay to determine the effect of plant extracts on bacterial viability and the cytotoxicity of the extracts in murine fibroblasts in culture. The results demonstrated significant antimicrobial activity, making the extract a potential candidate for the development of new antimicrobial agents. Furthermore, the non-toxic effects on fibroblasts highlight its potential to promote wound healing, suggesting that *Verbesina persicifolia* could serve as a natural alternative for therapeutic applications in wound management.

Keywords: *Verbesina persicifolia* · methanolic extract · antimicrobial properties · cytotoxicity · medicinal plants

1 Introduction

The *Verbesina* genus, belonging to the Asteraceae family, is one of the most extensive botanical groups in America, with more than 300 species distributed from western Canada to central Argentina. These species include herbs, shrubs and trees, characterized by the presence of bright yellow flowers and broad adaptability to different ecosystems [1, 2]. Among the pharmacological properties reported for various species of *Verbesina* include healing, antihypertensive, hypoglycemic, antitussive, anti-inflammatory and antimicrobial activities, in addition to its potential in the treatment

S. Han—The authors contributed equally to this work.

of diseases such as Chagas disease and prostate cancer [3–6]. The phytochemical richness of this genus includes bioactive compounds such as saponins, flavonoids, essential oils, tannins, phenolic compounds, sesquiterpenes, anthraquinones, cadinans, germacranes, monoterpenes, triterpenes, acetylenic compounds and alkaloids, which suggests its therapeutic potential in various pathologies [2, 5, 7–10]. Within this genus, *Verbesina persicifolia* is distributed in various regions of Mexico and is a species adapted to warm and semi-warm climates [11]. In traditional Mexican medicine, this species is widely used due to its medicinal properties, with ethnobotanical reports indicating its use in the treatment of gastritis, diabetes, wounds, kidney diseases, cancer, dysentery and dermatological conditions. The preparation of infusions and decoctions of its leaves and roots is a common practice to treat these conditions [11–13]. Recently, the anti-inflammatory potential of *g* has been explored through the development of a complex of zinc oxide nanoparticles coupled to methanolic extracts of the plant. In these studies, a significant anti-inflammatory effect was observed in vitro models, as well as a dose-dependent response in vivo models, highlighting its applicability in the treatment of inflammatory disorders and chronic diseases [14]. Despite the advances in research in this species, there is still a lack of scientific studies that allow us to rule out or validate the biological and toxicological effects, with the aim of generating new treatments. Within the *Verbesina* genus, antimicrobial activities have been reported. In this sense, the antimicrobial activity of *Verbesina persicifolia* extracts continues to be an insufficiently explored area of research, on the other hand, its toxicological activity is also explored to determine its possibility as a pharmacological treatment. Given the growing interest in the search for new antimicrobial agents derived from natural products, highlighting the need to generate new antimicrobial agents due to the growing resistance to antibiotics, the present study aims to evaluate the antimicrobial potential of the methanolic extract of *Verbesina persicifolia* leaves.

2 Materials and Methods

1. Collection of Plant Material

The leaves of *Verbesina persicifolia* were collected during the period September to October 2024 in the municipality of Martínez de la Torre, Veracruz, located at coordinates 20°08'43.0"N 97°03'56.1"W. Before drying the plant material, any leaves that were damaged, bitten by animals or yellowed were discarded. To dry the plant material, the previously selected leaves were spread on a bed of absorbent material and kept ventilated, away from direct sunlight, and stirred once a day to avoid excess moisture. In the end, the dry material was crushed and stored in a vacuum until its spindle. It should be noted that, for previous work, *Verbesina persicifolia* was collected at this site, in addition, flowering specimens were taken to the herbarium of the National Institute of Ecology (INECOL) for authentication (Boucher *Verbesina persicifolia* D.C.- AAM-002-XAL).

2. Preparation of the methanolic extract of *V. Persicifolia*

350 g of conditioned plant material were divided into two amber gallon containers and maceration began for 21 days with 3 liters of methanol. Every 7 days the extract was

concentrated with the help of a BUCHI-R210 rotary evaporator (BUCHI Corporation, New Castle, USA) with reduced pressure. It was placed in an oven at 40 °C to ensure the elimination of traces of methanol. Once the dry extract was obtained, it was stored in a refrigerator at 4 °C until used for microbiological tests.

For the preparation of the stock solution of *Verbesina persicifolia* extracts, 52 mg of powdered extract was dissolved in a sterile bottle, to which 104 µL of pure DMSO was added. The final volume was made up to 5.2 mL by adding 5.096 mL of sterile water, resulting in a final concentration of 10 mg/mL. From this stock solution, six different concentrations (5, 2.5, 1.25, 0.625, 0.3125 and 0.15625 mg/mL) were prepared for use in the Minimum Inhibitory Concentration (MIC) and AlamarBlue (cell viability) assays.

3. Preparation of microbiological cultures

Cultures of *Pseudomonas aeruginosa* (ATCC 10145), *Enterococcus faecalis* (ATCC 29212), *Staphylococcus aureus* (ATCC 25923) (ATCC, Manassas, VA) were maintained on Difco nutritious agar and Difco nutritious broth (Becton, Dickinson, and Company [BD], New Jersey, United States) as a liquid phase and were incubated under conditions aerobic at 37 °C for 48 hours in a microbiological incubator (Heratherm™ Thermo Scientific, Waltham, MA, USA). The viability of the cultures was determined by phase contrast microscopy on a CX43 microscope (Olympus Corp., Tokyo, Japan) 24 hours after culture, and the purity of the cultures was checked by Gram staining. The cells were passed every 48 hours by adding 1 ml of bacteria from the culture in fresh medium. Incubation was performed as described above. On test days, liquid cultures were centrifuged at 1000 rpm for 10 min at 4°C and the resulting pellet was resuspended at the appropriate concentration for each experiment described below.

4. Antimicrobial activity assay

The antimicrobial activity was evaluated by the Minimum Inhibitory Concentration (MIC) assay in 96-well plates, which was carried out as follows: 100 µL of each concentration of the *Verbesina persicifolia* extracts was added to the 96-well plate. 100 µL of bacterial inoculum (*Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Enterococcus faecalis*) was added to each well at a concentration on the McFarland scale of 3 corresponding to 6×10^8 cells. The positive control corresponded to a well inoculated only with the corresponding bacteria. The negative control was prepared with the test extract only. The plate was incubated for 24 hours at 37°C in a microbiological incubator (Heratherm™ Thermo Scientific, Waltham, MA, USA).

Subsequently, 100 µL from each well were added to a new plate. After incubation, absorbance was measured at 620 nm using a Multiskan™ FC Microplate Photometer (Thermo Scientific, Waltham, MA, USA). A wavelength of 620 nm was selected because it provides accurate measurements of bacterial viability, as it minimizes interference from medium components and other metabolites in the visible spectrum. Additionally, this wavelength optimizes the signal-to-noise ratio, avoiding unwanted absorption at other spectral ranges, such as 400 nm or 700 nm. Experiments were performed in triplicate and results are shown as means $\pm$ standard error. The "n" of this study was 3, since that is the number of replicates per experimental condition that were evaluated.

5. Bacterial Cell Viability Assay with AlamarBlue

For cell viability assay with AlamarBlue, 200 μ L of each concentration of *Verbesina persicifolia* extracts was added to the 96-well plate, followed by 20 μ L of corresponding bacterial inoculum. The plate was incubated for 24 hours at 37°C. Subsequently, 10 μ L of AlamarBlue reagent was added to each well, and the plate was incubated for an additional 4 hours at 47°C. Finally, absorbance was measured at 620 nm using the Multiskan FC microplate reader (Thermo Fisher Scientific, MA, USA). Experiments were performed in triplicate and results are shown as means $\pm$ standard error.

6. Cytotoxicity assay

A cytotoxicity assay of the methanolic extracts of *V. persicifolia* was carried out at concentrations of 0.1, 0.5, 1, 1.5 and 2 mg/mL in cultured fibroblasts. Initially, the mouse embryonic fibroblast cell line NIH-3T3 (ATCC CRL-1658) was cultured in MEM medium (Gibco, Grand Island, NY, USA.) enriched with 10% fetal bovine serum (Gibco, Grand Island, NY, USA.) and 1% penicillin-streptomycin (Gibco, Grand Island, NY, USA) in 96-well plates at a density of 2000 cells per well. The plates were incubated for 2 days at 37°C in a humidified incubator (Forma Steri-Cycle i160, Thermo Fisher Scientific, MA, USA) under the following conditions: 95% air and 5% CO₂. The morphology, confluency, and characteristics of the cultures were visualized on an Olympus IX53 inverted microscope (Olympus Corp., Tokyo, Japan). Treatments with *V. persicifolia* extracts began once 80% cell confluency was reached. For this, the cells were washed twice using 100 μ L of sterile PBS IX, subsequently 100 μ L of each of the *V. Persifolia* solutions in MEM medium were added to each test well. Cell viability was analyzed using the AlamarBlue™ assay (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA.). For this, 100 μ L of MEM medium supplemented with 10% AlamarBlue was added to each well in treated 96-well plates and the plates were incubated for 4 h in a humid atmosphere at 37 °C and 5% CO₂. Subsequently, the absorbance of the samples was determined at 570 nm using a Multiskan™ FC microplate spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Experiments were performed in triplicate and results are shown as means $\pm$ standard error.

7. Data Analysis

All experiments were performed in triplicate. The results were expressed as the mean $\pm$ standard error. For inferential analysis of the data, the ANOVA test was used to compare between groups. The value of $p < 0.05$ was considered statistically significant. Data analysis was performed using SigmaPlot Version 16.0 software (SPSS Inc., Chicago, IL, USA).

3 Results and Discussion

The results of the minimum inhibitory concentration (MIC) test are presented in Table 1, where the absorbance values obtained for the antimicrobial activity of the methanolic extract of *Verbesina persicifolia* against *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Enterococcus faecalis* are shown. Different concentrations of the extract were

analyzed, from 5.00 mg/mL to 0.15 mg/mL. A progressive decrease in absorbance values was observed as the extract concentration increased, suggesting a dose-dependent inhibition of bacterial growth. The untreated control had the highest absorbance in all microorganisms, indicating unrestricted bacterial growth (Fig. 1).

In the plates inoculated with *S. aureus*, the absorbance gradually decreased until reaching a value of 0.163 to 0.15 mg/mL, evidencing a sustained antimicrobial effect. *P. aeruginosa* showed a more pronounced reduction, with an absorbance value of 0.742 at 5.00 mg/mL, indicating a greater sensitivity to the extract. In the plates inoculated with *S. aureus*, the absorbance gradually decreased until reaching a value of 0.163 to 0.15 mg/mL, evidencing a sustained antimicrobial effect. *P. aeruginosa* showed a more pronounced reduction, with an absorbance value of 0.742 at 5.00 mg/mL, indicating a greater sensitivity to the extract.

Table 1. Absorbance values obtained from the MIC test.

<i>S. aureus</i>			<i>P. aeruginosa</i>			<i>E. faecalis</i>		
Concentration			Concentration			Concentration		
(mg/mL)	Average*	SE	(mg/mL)	Average*	SE	(mg/mL)	Average*	SE
Control	1.381	0.033	Control	1.413	0.058	Control	1.149	0.090
5.00	1.111	0.074	5.00	0.742	0.094	5.00	0.928	0.060
2.50	0.810	0.015	2.50	0.736	0.017	2.50	0.638	0.088
1.25	0.456	0.094	1.25	0.398	0.020	1.25	0.598	0.019
0.62	0.475	0.017	0.62	0.333	0.036	0.62	0.557	0.052
0.31	0.235	0.055	0.31	0.212	0.052	0.31	0.516	0.052
0.15	0.163	0.034	0.15	0.108	0.037	0.15	0.497	0.045

* The experiments were performed in triplicate.

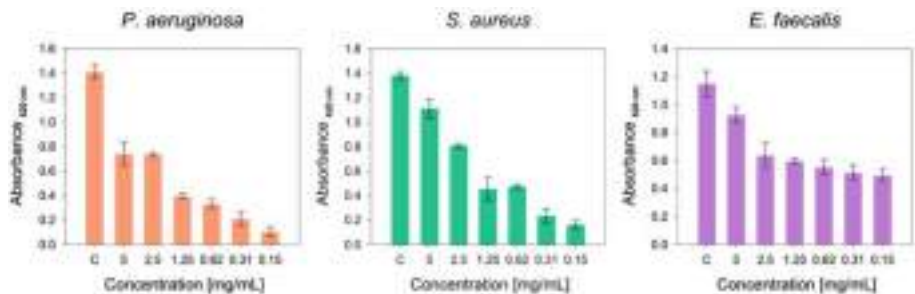


Fig. 1. Determination of the minimum inhibitory concentration of the methanolic extract of *V. Persifolia* against *Pseudomona aeruginosa*, *Staphylococcus aureus* and *Enterococcus faecalis*, after 24 h of treatment. The experiments were performed in triplicate. Graphs represent means and standard errors (SE).

The cell viability test on different bacterial strains using the AlamarBlue test confirmed the antimicrobial activity of the methanolic extract. A decrease in the percentage of bacterial survival was observed as the concentration of the extract increased. In *S. aureus*, bacterial survival decreased from 100% in the control to 29.4% at 5 mg/mL. *P. aeruginosa* showed the highest sensitivity, with a reduction from 100% to 11.6% at the same concentration. In contrast, *E. faecalis* exhibited higher resistance, with survival of 41.4% at 5 mg/mL. Greater variability in the standard error values was observed at high concentrations, suggesting a greater dispersion in the data under these conditions. These results confirm the differential antimicrobial activity of the extract, with *P. aeruginosa* being the most susceptible microorganism and *E. faecalis* the most resistant (Table 2; Fig. 2).

Table 2. Determination of bacterial cell viability against the methanolic extract of *V. persifolia*.

<i>P</i>								
<i>S aureus</i>	% Survival	SE	<i>aeruginosa</i>	% Survival	SE	<i>E. faecalis</i>	% Survival	SE
Control	100	2.947	Control	100	2.172	Control	100	4.9639
0.15 mg/ml	75.9	1.437	0.15 mg/ml	73.3	6.342	0.15 mg/ml	88.9	3.7076
0.31 mg/ml	69.0	2.841	0.31 mg/ml	60.2	3.593	0.31 mg/ml	58.5	4.8135
0.62 mg/ml	65.8	4.020	0.62 mg/ml	55.8	7.248	0.62 mg/ml	49.8	2.7699
1.25 mg/ml	41.1	3.396	1.25 mg/ml	42.1	4.432	1.25 mg/ml	48.6	6.4148
2.5 mg/ml	49.4	2.850	2.5 mg/ml	28.5	3.933	2.5 mg/ml	48.3	1.9375
5 mg/ml	29.4	1.397	5 mg/ml	11.6	4.308	5 mg/ml	41.4	4.0411

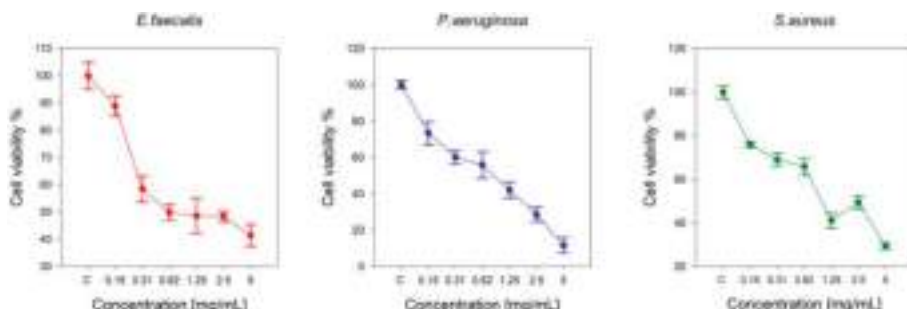


Fig. 2. Results of the cell viability test in bacterial cultures treated with different concentrations of the methanolic extract of *V. persifolia*. The experiments were performed in triplicate. The graph shows the mean values of the percentage of cell survival with the corresponding standard errors.

Finally, cytotoxicity assays in NIH 3T3 cells revealed that the methanolic extract presents a cytotoxic effect at concentrations higher than 1 mg/mL. However, concentrations of 0.1, 0.5 and 1 mg/mL did not significantly affect cell viability ($p>0.05$), suggesting that these conditions could be safe for therapeutic applications (Fig. 3).

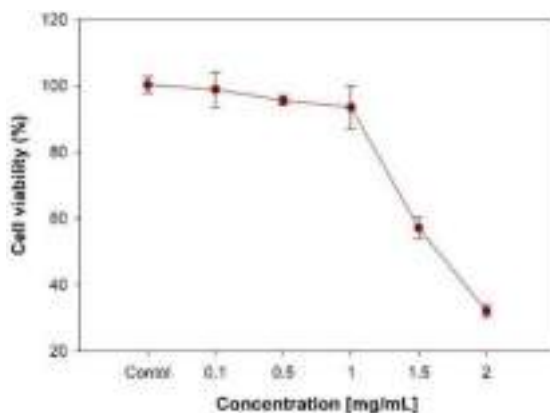


Fig. 3. Cytotoxicity assay results in NIH3T3 fibroblasts treated with different concentrations of *V. persicifolia*. Experiments were conducted in triplicate, and the data are presented as means $\pm$ standard error.

These findings support the biomedical potential of *V. persicifolia* extract, suggesting its application in the design of antimicrobial therapies and wound healing strategies with an adequate safety margin for use in biological systems.

4 Conclusion

The present investigation demonstrates that the culture of NIH 3T3 cells was successful, particularly when the MEM medium was used, which favored cell viability. Likewise, the tests carried out with the methanolic extracts of *Verbesina persicifolia* confirmed its biological activity. After 24 hours of treatment, preliminary results indicated that these extracts do not significantly affect cell proliferation at safe concentrations, supporting their potential as therapeutic agents with a favorable safety profile. Furthermore, moderate antimicrobial activity was evident against the microorganisms evaluated, which suggests its applicability in the development of biomedical devices with antimicrobial properties aimed at wound healing. The relevance of these findings lies in the optimization of conditions for the use of NIH 3T3 cells in subsequent studies and in the generation of preliminary evidence on the safety of *V. persicifolia* extracts. These results open new opportunities for research of this plant in different biomedical applications, including its possible use in anticonvulsant therapies and other medical treatments. It is crucial to continue with studies in cellular and animal models to further evaluate its biological effects, as well as explore its combination with other therapeutic compounds to enhance its clinical efficacy.

References

1. Mora FD, Alpan L, McCracken VJ, Nieto M (2013) Chemical and biological aspects of the genus *Verbesina*. *Nat Prod J* 3(2):140–150
2. Ramseyer J et al (2017) Eudesmane sesquiterpenes from *Verbesina lanata* with inhibitory activity against grapevine downy mildew. *J Nat Prod* 80(12):3296–3304
3. Delle Monache G et al (1992) Caracasamide, a novel hypotensive agent from *Verbesina caracasana*. *Bioorg Med Chem Lett* 2(5):415–418. [https://doi.org/10.1016/S0960-894X\(00\)80159-6](https://doi.org/10.1016/S0960-894X(00)80159-6)
4. Dalla Via L, Mejia M, García-Argáez AN, Braga A, Toninello A, Martínez-Vázquez M (2015) Anti-inflammatory and antiproliferative evaluation of 4 β -cinnamoyloxy, 1 β , 3 α -dihydroxyeudesm-7, 8-ene from *Verbesina persicifolia* and derivatives. *Bioorg Med Chem* 23(17):5816–5828. <https://doi.org/10.1016/j.bmc.2015.07.002>
5. Arciniegas A, Pérez-Castorena AL, Villaseñor JL, de Vivar AR (2020) Cadinenes and other metabolites from *Verbesina sphaerocephala*: a Gray. *Biochem Syst Ecol* 93:104183. <https://doi.org/10.1016/j.bse.2020.104183>
6. García-Bores AM et al (2020) *Verbesina crocata*: a pharmacognostic study for the treatment of wound healing. *Saudi J Biol Sci* 27(11):3113–3124. <https://doi.org/10.1016/j.sjbs.2020.08.038>
7. Compagnone RS, Bermudez J, Ibáñez G, Diaz B, Garrido MR, Israel A, Suarez AI (2008) New guanidine alkaloids from the leaves of *Verbesina peraffinis*. *Nat Prod Commun* 3(4):1934578X0800300407. <https://doi.org/10.1177/1934578X0800300407>
8. Jakupovic J, Ellmauerer E, Jia Y, Bohlmann F, Domínguez XA, Hirschmann GS (1987) Further eudesmane derivatives from *Verbesina* species. *Planta Med* 53(01):39–42. <https://doi.org/10.1055/s-2006-962614>
9. Verma V, Chaudhary M, Srivastava N (2019) Antioxidative properties of isolated saponins of *Verbesina encelioides* (Cav.) Benth. & Hook. fil ex Gray and SEM studies of synthesized green nanoparticles for acne management. *Plant Sci Today* 6(1):575–582. <https://doi.org/10.14719/pst.2019.6.sp1.675>
10. - Jain SC, Renu Singh RS, Renuka Jain RJ (2008) Antimicrobial and antioxidant potentials of *Verbesina encelioides* (Cav.) Benth. and Hook. Fil ex Gray
11. Fonseca-Chávez RE, Rivera-Levario LA, Vázquez-García L (2020) Guía ilustrada de plantas medicinales en el Valle de México. México: Instituto Nacional de los Pueblos Indígenas
12. Alfaro MÁM, Oliva VE, Cruz MM, García GM, Olazcoaga GT, León AW (1995) Catálogo de plantas útiles de la Sierra Norte de Puebla. Universidad Nacional Autónoma de México, México
13. Domínguez-Barradas C, Cruz-Morales GE, González-Gándara C (2015). Plantas de uso medicinal de la Reserva Ecológica “Sierra de Otontepec”. municipio de Chontla, Veracruz, México. *CienciaUAT* 9(2):41–52
14. Soto-Ojeda GA, Ramos-Santos E, Gómez-Sedeño JM, Villanueva-Lendechy MA (2022) Evaluación de la administración crónica de ZnO verbesina persi- cifolia ZnO-Vp sobre la actividad cicatrizante e hipoglucemiante en ratas hiperglucémicas. *Revista ingeniantes*. 9(3):16–20



Influence of Liner Material Poisson's Ratio on Contact Pressure Distribution in Transtibial Prosthetic Sockets: A Finite Element Analysis Study

Agata Mrozek-Czajkowska^(✉) 

Poznan University of Technology, Poznan, Poland
agata.mrozek-czajkowska@put.poznan.pl

Abstract. Additive manufacturing has significantly advanced the field of prosthetics, allowing for the production of highly customized prosthetic components. Among these, the prosthetic socket plays a critical role in securely attaching the residual limb to the prosthesis. The interface pressure between the residual limb and the prosthetic socket considerably impacts user comfort and quality of life. This study explores the influence of liner material's Poisson's ratio on contact pressure distribution within transtibial prosthetic sockets using Finite Element Analysis (FEA). Auxetic metamaterials, characterized by a negative Poisson's ratio, exhibit unique mechanical properties that enhance adaptability and comfort in prosthetic applications. By analyzing the pressure distribution for various Poisson's ratio values, the study reveals that auxetic materials contribute to a more uniform pressure distribution, reducing peak stresses on the residual limb. The findings suggest that integrating auxetic materials into prosthetic socket design can significantly improve the comfort and performance of prosthetic users. Further research should investigate additional mechanical properties to optimize prosthetic liner materials.

Keywords: Prosthetics · FEA · Auxetics

1 Introduction

Additive manufacturing has significantly expanded its applications in the field of prosthetics, enabling the production of prosthetic components tailored to the individual anatomical characteristics of patients. Among these components, the prosthetic socket plays a crucial role in securely attaching the residual limb to the prosthesis. The interface pressure between the residual limb and the prosthetic socket is a key factor affecting the user's comfort and overall quality of life. Due to the unique shape of each residual limb, a high level of customization is required in socket design. Finite Element Analysis (FEA) provides a valuable tool for assessing interface pressure distribution and optimizing socket geometry, which improves the prosthetic design process [1, 2].

The assessment of patient comfort is typically based on numerical criteria that define the pain threshold and the patient's pain tolerance. In the literature, various criteria are used to determine patient comfort, including pressure criteria in the residual limb volume for different areas [3–7], deformation criteria [8, 9], and critical stress criteria [10].

Auxetic metamaterials, characterized by a negative Poisson's ratio, exhibit an unconventional deformation behavior that differs from traditional materials. While conventional materials expand laterally under compression and contract when stretched, auxetic materials behave oppositely—contracting transversely under compression and expanding when stretched. This unique mechanical response was first described by Evans in 1991 and has since been the subject of extensive research [11, 12]. The exceptional properties of auxetic structures, including high energy absorption and impact resistance, have led to their application in protective equipment such as helmets and ballistic barriers, [13, 14].

Recent studies have explored the potential of auxetic materials in medical applications, particularly in prosthetics. The adaptability of auxetic structures allows for improved adjustment to anatomical shapes, making them an attractive option for prosthetic sockets, where precise fitting is essential [15]. By integrating auxetic metamaterials into prosthetic socket design, it is possible to enhance user comfort and reduce pressure concentrations on the residual limb. Researchers have highlighted the benefits of auxetic materials in optimizing prosthetic components, particularly in areas where pressure distribution plays a critical role [16–18]. Additionally, investigations into the incorporation of auxetic structures in prosthetic foot designs suggest that these materials can effectively replicate the biomechanical functions of the human foot [19, 20]. Although auxetic metamaterials demonstrate significant potential for enhancing prosthetic devices, their clinical implementation remains in the developmental stage.

This study aims to evaluate the influence of liner material Poisson's ratio on contact pressure distribution in transtibial prosthetic sockets using FEA. By analyzing different liner materials, the effect of variations in Poisson's ratio on the mechanical interaction between the socket and residual limb was investigated, ultimately contributing to improved comfort and performance for prosthetic users.

2 Methods

The finite element model consists of seven components: soft tissues of the residual limb, four bones (patella, fibula, tibia, femur), liner and prosthetic socket (Fig. 1). The skin and muscle tissue were treated as homogenized soft tissues. The models of the residual limb and bone were obtained from the OpenLimb project [21]. Based on this geometry, the socket and liner were modeled using nTop software (version 5.9.2).

The hyperelastic behavior of the soft tissues in the residual limb was defined using the generalized Mooney-Rivlin solid strain energy function [22]. Bones and the prosthetic socket were modeled as rigid bodies. The liner was represented using the neo-Hookean model, with an applied Young's modulus of 10 MPa. The Poisson's ratio was varied between 0.3 and 0.45, including corresponding negative values. Metamaterials are artificially structured materials, allowing for obtaining a wide range of Poisson's ratio values. The selected values were chosen based on materials commonly used in the

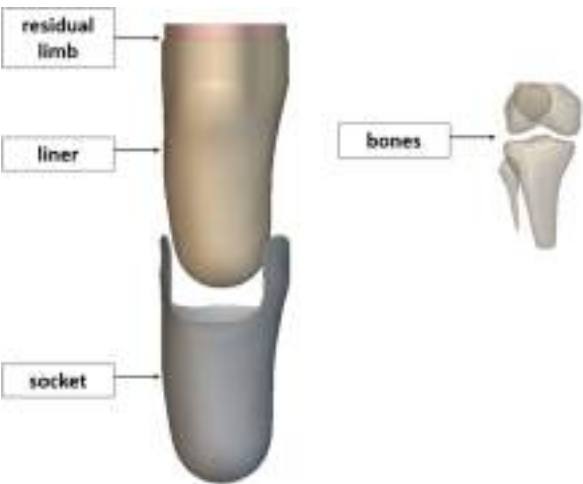


Fig. 1. Components included in the finite element model: prosthetic socket, liner, soft tissues of the residuals limb and four bones: patella, fibula, tibia, femur.

manufacturing of prosthetic liners. This selection was made to emphasize the influence of the Poisson’s ratio on the mechanical behavior of these materials, ensuring the study’s relevance to practical applications in prosthetic design (Fig. 2).

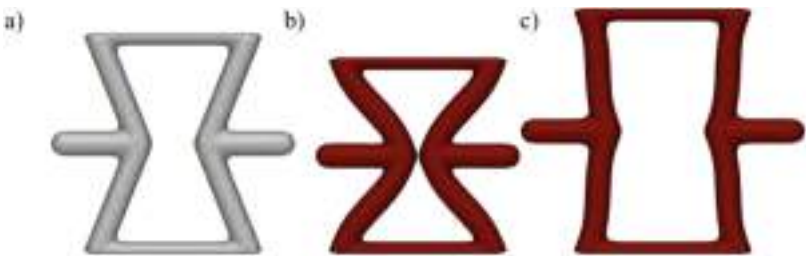


Fig. 2. Example of an auxetic unit cell: a) base geometry, b) under compression, c) under tension.

In this study, the donning procedure of press-fit sockets was simulated. During the analysis, a displacement vector was applied to the socket to ensure proper positioning. The upper surface was fixed. A tied contact condition was applied between the bones and the residual limb, as well as between the liner and the residual limb (Fig. 3). A sliding contact condition was assigned between the socket and the liner. The simulation was conducted using FEBio software (version 2.6.0). Contact pressure was measured on the surface of the residual limb. The finite element mesh consisted of 10-node quadratic tetrahedral element (689 452 elements).

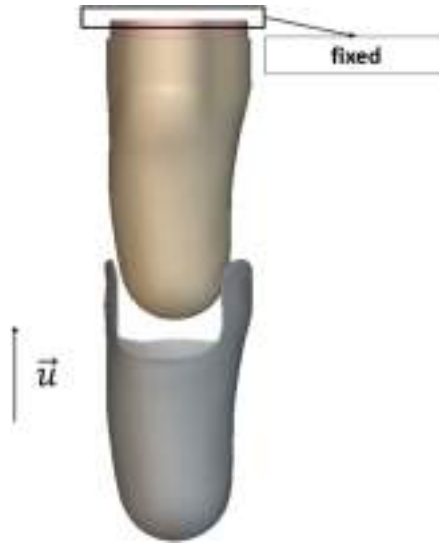


Fig. 3. Applied boundary conditions.

3 Results and Discussion

Table 1 presents the obtained values of contact pressure during the donning procedure, expressed in megapascals (MPa), for different Poisson's ratio values of the liner material. The results highlight a significant difference in contact pressure distribution between positive and negative Poisson's ratio values.

For positive Poisson's ratio values (0.30 – 0.45), the contact pressure varies between 1.32 MPa and 0.52 MPa. The highest contact pressure (1.32 MPa) is observed at 0.30 Poisson's ratio, while the lowest (0.52 MPa) occurs at 0.40 Poisson's ratio. These fluctuations suggest that traditional liner materials with a positive Poisson's ratio influence pressure distribution, potentially leading to higher localized stresses (Table 1).

Conversely, for negative Poisson's ratio values, the contact pressure remains relatively stable, ranging between 0.29 MPa and 0.31 MPa, regardless of the specific Poisson's ratio value. This indicates that auxetic materials (negative Poisson's ratio) contribute to a more uniform pressure distribution, potentially enhancing user comfort by reducing peak pressures and strains on the residual limb (Fig. 4).

These findings suggest that utilizing a liner with a negative Poisson's ratio could be beneficial in prosthetic socket design, as it may help in minimizing pressure concentrations and improving overall comfort for prosthetic users. Further investigations could explore the impact of additional mechanical properties, such as material stiffness and thickness, on pressure distribution.

The finite element model utilizes geometry derived from the OpenLimb project. Nevertheless, a patient-specific geometric representation would enhance accuracy by capturing individual anatomical features more precisely. Moreover, adjusting material model parameters to align with specific patient characteristics would improve the reliability of the simulations, enabling a more realistic assessment of mechanical behavior.

Table 1. Obtained value of contact pressure during donning procedure-expressed in [MPa]

Poisson’s ratio	Positive	Negative
0.30	1.32	0.31
0.35	0.70	0.31
0.40	0.52	0.29
0.45	0.71	0.29

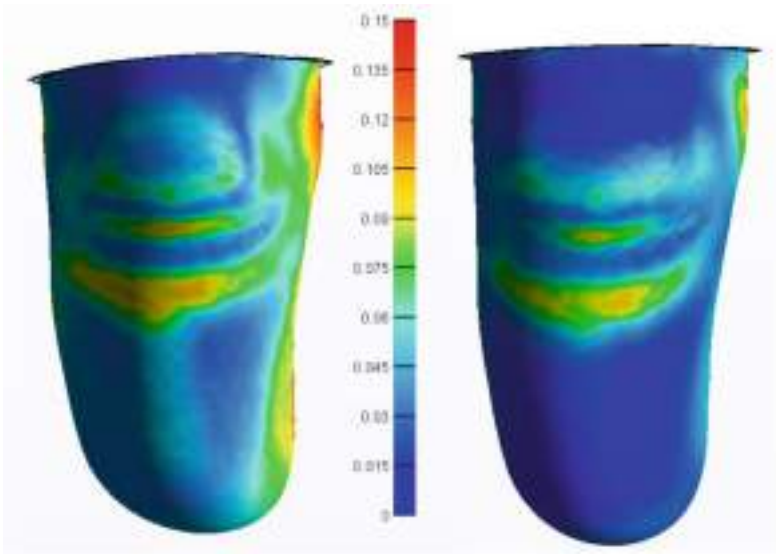


Fig. 4. Comparison of obtained strain for different values of Poisson’s ratio (0.3 on the left, -0.3 on the right).

In the current model, soft tissue is treated as a homogenized material that includes muscle, adipose tissue, and skin – distinguishing these tissue types would be essential for enhancing the model’s physiological relevance and predictive capability.

4 Conclusions

This study highlights the potential benefits of using auxetic metamaterials in transtibial prosthetic socket design. Through Finite Element Analysis, we observed that traditional liner materials with a positive Poisson’s ratio lead to variations in contact pressure, potentially causing localized stress concentrations. In contrast, liners with a negative Poisson’s ratio demonstrated a more uniform pressure distribution, reducing peak pressures on the residual limb. These findings suggest that incorporating auxetic materials in prosthetic socket design can enhance user comfort and mitigate pressure-related complications.

The article considers the strains that occur during and after the donning procedure. Since human movement is inherently dynamic and not limited to vertical forces, lateral

and multi-directional pressures will be taken into account. Future research will also expand the analysis to examine the impact of pressure during various phases of the gait cycle. The results reinforce the importance of advanced material selection in improving prosthetic performance and user experience.

Acknowledgements. This research was funded by the grant 0612/SBAD/3642, Funded by the Ministry of Science and Higher Education, Poland. Part of the studies were realized in scope of the grant entitled “Biomedical Innovations through Digital Transformation of Additive Technologies and Knowledge Exchange – BIOMEDIX” supported by ERASMUS+ (2024-1-LV01-KA220-HED-000255929).

References

1. Steer JW et al (2021) *Prosthet Orthot Int* 45(2):138–146
2. Lacroix D et al (2011) *Ann Biomed Eng* 39:2972
3. Lee WC, Zhang M, Mak AF (2005) Regional differences in pain threshold and tolerance of the transtibial residual limb: including the effects of age and interface material. *Arch Phys Med Rehabil* 86(4):641–649. <https://doi.org/10.1016/j.apmr.2004.08.005>
4. Zhang M, Lee WCC (2006) Quantifying the regional load-bearing ability of transtibial stump. *Prosthet Orthot Int* 30(1):25–34
5. Finocchietti S, Nielsen M, Mørch CD, Arendt-Nielsen L, Graven-Nielsen T (2011) Pressure-induced muscle pain and tissue biomechanics: a computational and experimental study. *Eur J Pain* 15(1):36–44
6. Safari MR, Tafti N, Aminian G (2014) Socket interface pressure and amputee reported outcomes for comfortable and uncomfortable conditions of patellar tendon bearing socket: a pilot study. *Assist Technol* 27(1):24–31. <https://doi.org/10.1080/10400435.2014.949016>
7. Ghoseiri K, Rastkhadiv MY, Allami M (2018) Evaluation of localized pain in the transtibial residual limb. *Can Prosthet Orthot J*. <https://doi.org/10.33137/cpoj.v1i2.32028>
8. Gefen A, van Nierop B, Bader DL, Oomens CW (2008) Strain-time cell-death threshold for skeletal muscle in a tissue-engineered model system for deep tissue injury. *J Biomech* 41(9):2003–2012. <https://doi.org/10.1016/j.jbiomech.2008.03.039>
9. Portnoy S, Siev-Ner I, Shabshin N, Kristal A, Yizhar Z, Gefen A (2009) Patient-specific analyses of deep tissue loads post transtibial amputation in residual limbs of multiple prosthetic users. *J Biomech* 42(16):2686–2693. <https://doi.org/10.1016/j.jbiomech.2009.08.019>
10. Linder-Ganz E, Engelberg S, Scheinowitz M, Gefen A (2006) Pressure-time cell death threshold for albino rat skeletal muscles as related to pressure sore biomechanics. *J Biomech* 39(14):2725–2732. <https://doi.org/10.1016/j.jbiomech.2005.08.010>
11. Lim, T.-C., *Mechanics of Metamaterials with Negative Parameters*. Singapore: Springer Singapore, 2020. <https://doi.org/10.1007/978-981-15-6446-8>
12. Lim T-C (2015) *Auxetic Materials and Structures*. Springer, Singapore. <https://doi.org/10.1007/978-981-287-275-3>
13. Krishnan BR, Biswas AN, Kumar KVA, Sreekanth PSR (2022) Auxetic structure metamaterial for crash safety of sports helmet. *MaterToday Proc* 56:1043–1049
14. Tahir D, Zhang M, Hu H (2022) Auxetic materials for personal protection: a review. *Phys Status Solidi (B)* 259:2200324
15. Fardan MF, Lenggana BW, Ubaidillah U, Choi SB, Susilo DD, Khan SZ (2023) Revolutionizing prosthetic design with auxetic metamaterials and structures: a review of mechanical properties and limitations. *Micromachines* 14:1165

16. Devin KM, Tang J, Hamilton AR, Moser D, Jiang L (2024) Assessment of 3D printed mechanical metamaterials for prosthetic liners. *Proc Inst Mech Eng Part H J Eng Med* 238:348–357
17. Brown N, Owen MK, DesJardins JD, Garland A, Fadel GM (2020) Metamaterial design for targeted limb-socket interface pressure offloading in transtibial amputees. In: *Proceedings of the ASME Design Engineering Technical Conference, Virtual*, 17–19 August 2020. American Society of Mechanical Engineers (ASME): New York
18. Kowalczyk M, Jopek H (2020) Numerical analysis of the lower limb prosthesis subjected to various load conditions. *Vib Phys Syst* 31:2020309
19. Um HJ, Kim HS, Hong W, Kim HS, Hur P (2021) Design of 3D printable prosthetic foot to implement nonlinear stiffness behavior of human toe joint based on finite element analysis. *Sci Rep* 11:19780
20. Kim HS, Um HJ, Hong W, Kim HS, Hur P (2021) Structural design for energy absorption during heel strike using the auxetic structure in the heel part of the prosthetic foot. In: *Proceedings of the 2021 18th International Conference on Ubiquitous Robots (UR)*, Gangneung, Republic of Korea, 12–14 July 2021
21. Sunderland FE, Bramley JL, Al-Dirini RMA, Steer JW, Worsley PR, Dickinson AS (2023) OpenLimb: an open source transtibial residual limb model for simulation and design. *Prosthetics Orthotics Int* 47, 165. <https://doi.org/10.1097/pxr.0000000000000240>. <https://github.com/abel-research/openlimb>. Accessed 31 Jan 2025
22. Portnoy S et al (2009) *Ann Biomed Eng* 37:2583

**Biomeasurements, Biomedical Sensors,
Health Technology Assessment,
and Rehabilitation Engineering**



User Expectations on Wearable Medical Devices in Latvia

Nadezhda Semjonova¹ , Roberts Malnieks¹, and Alexei Katashev² 

¹ Economics and Business Institute, Riga Technical University, Kipsalas Street 6B, Riga 1048, Latvia

Nadezhda.Semjonova@rtu.lv

² Mechanical and Biomedical Engineering Institute, Riga Technical University, Kalnciema Street 6, Riga 1048, Latvia

Abstract. This study explores expectations of potential users of wearable medical devices in Latvia. The opinions of healthy individuals were collected using on-line free survey portal, this limited the study to the population of active internet and IT users. Out of 89 respondents, 44 already use wearable devices, like smart watches of fitness activity meters. Latvian respondents consider price as the most important factor, determining intention to purchase the device. This is different from other developed countries, where easiness to use is perceived as the most important factor. Alongside, Latvian respondents perceive wearable devices as mostly suitable for fitness activity monitoring and seldom see any health benefits except of fitness tracking. The medical doctors could change this opinion, as the Latvian users consider medical doctors as the most creditable information sources about wearable medical devices.

Keywords: Wearable devices · Fitness tracking · Health monitoring · User expectations · Latvia

1 Introduction

Continuous progress in medical technologies, prolonged human lifespan, changes in the age structure of society, unhealthy lifestyle and eating habits, occupational stresses – all this pose challenges to modern healthcare. The remedy may be the remote, homecare, and personalized solutions, that could drastically reduce hospital load by transferring healthcare toward patient home.

Wearable medical devices are considered as one of the most promising approaches for personalized healthcare monitoring and the prerequisite for the development of a more cost-effective and sustainable healthcare system. The rich literature on wearable medical devices covers reviews of academic advances [1, 2], comparison of commercially available products [3], case studies and success stories [4, 5].

Despite exciting perspectives, the regular use of wearable devices is still limited to activity monitors and fitness trackers [6]. People stop using their devices even though they promise notorious health benefits [7]. Such limitations could be related to how users perceive usability of the device.

More than ten years ago Money *et.al.* mentioned that manufacturers seldom consider patients as a source of valuable input for the design of medical devices [8]. Although the patients' advice and feedback are important factors to meet user expectations [9], innovators still do not involve patients in development of neither medical devices [10], nor home care devices [11].

The discrepancy between intensive research and rather humble commercial implementations fostered studies of user expectation and experience while using wearable devices. Ferguson et al. provided comprehensive review on research, targeted to explore the perceptions and experiences of older adults about using wearable cardiac monitoring technologies, and to identify barriers and facilitators of acceptance and uptake of these devices in clinical practice [12].

Kim et al. (2017) evaluated user experience in terms of ease of use, learnability (how easily novice users can master the device), wearability (how comfort the device is), utilitarian value (benefit user receives from the very function of the device), hedonic value (how fun is to use the device), and aesthetic value (how elegant is the device) [13].

Following Park, the intention to use smart wearable devices is determined by five factors: satisfaction, enjoyment, usefulness, flow state (level of user involvement), and cost [14]. Huarng, Hui-Kuang Yu and Lee mentioned data privacy, cost burden, and perceived ease of use as key factors. In both cases, the cost was the least important, or even unimportant factor [15].

Appreciation of user expectancies seems to be an important issue the developers and distributors should consider, but there is a question, whether these expectancies are similar in different countries. Over recent years, facing COVID-19 pandemic, Latvian people meet some economic difficulties. One could suspect cost issues could play a more significant role in Latvian customers' willingness to use wearable medical device. The purpose of the present paper is to assess the expectations of potential medical devices users in Latvia.

2 Materials and Methods

The research was designed as a cross - sectional study, based on the collection of public opinion using an anonymous questionnaire. The research group included both users and non-users of the wearable devices. The questionnaire consisted of 15 questions, including a field for narrative comments.

Three questions provided sociodemographic data: age, sex, and occupation. Two questions inquired about what type of device, if any, the responder uses, and how often.

Other questions collected responder's opinion on different aspects of use and functionality of wearable devices (see Table 1 in the Result section). Most of the questions offered pre-set limited choice response with an option to give custom narrative answer. Two questions used 5-point qualitative Likert scale, ranged from 1 (not important) to 5 (very important). Two questions (including free comment section) presumed narrative answer.

The size of the poll was chosen to provide confidence interval for estimates of proportions within $\pm 10\%$ at the confidence level 90%, roughly estimating Latvian

population as 1 900 000 [16]. The corresponding minimal sample size, equal to 64, was calculated using an online tool [17].

The questionnaires were disseminated and collected using on-line portal “VisiDati” [18]. The link to the survey was disseminated via authors’ social media, as well as via e-mails. Hereby, the study is limited to the part of the population that has access to the Internet and uses it. It is reasonable to presume that participants were from the same social group as the authors of research (presumably middle class). There were no limitations or exclusion criteria for respondents. One should note that the present study is not intended to provide unbiased estimation of the number of wearable device users, but to collect users’ opinions.

In total, 89 questionnaires were collected. This number of questionnaires provided an uncertainty of the proportion estimate $\pm 9\%$ with the confidence level 90%. This number was in good agreement with optimum (90), derived by Caron et al. for usability studies [20]. For the limited choice-type questions, the proportions of certain answers were calculated. For questions 7 and 10 the average Likert estimate was calculated for each of the features in question, and the results were used to order features by importance.

3 Results

Out of 89 respondents, 29 (33%) were males, 60 (67%) were females. The average age of responders was 36.7 years. Most respondents (62 people, approximately 70%) were employees. Another 30% of respondents were students, self – employed, entrepreneurs, unemployed etc.

These results demonstrate that respondents represented young economically active people. The way the questionnaires were disseminated implies, that respondents were active Internet and social networks users, skilled enough to not hesitate “go digital”, that is critical to benefit from the modern information technologies [20]. They do not have “fear of technologies”, typical for elder population [21].

Despite such “progressive” respondents, about half of them (45 individuals, or 50.6%) do not own any wearable device. Among 44 owners, the majority (30 subjects) own smartwatches, 12 individuals own fitness activity meters, one person owns smart glasses, and one person owns smart garment. Smart watches are the most popular device, that could be explained by the wider availability of such devices on the market.

At the same time, only 42 respondents out of 89 (47.2%) reported they do not use wearable devices at all. The respondents that use wearable devices but do not own them used fitness activity meters owned by fitness centers. Most users (44 out of 47, or 94% of all users) use their devices at least once per week, 38 respondents do it daily. Among all respondents, only 3 individuals (6%) use their devices once a month and less frequently. This may indicate that customers generally purchase a device when they feel the need for use of the device, although there is a risk that responders may overestimate how frequently they use their devices. One needs to note that the present study does not estimate the duration of continuous use of the device, as the purpose of the questions 4 and 5 of the Questionary was to characterize responder group in respect to the use of wearable device, not to explore behavior of the users.

The responses to the substantive questions are summarized in Table 1. While discussing the reason why people refuse to purchase and use of the wearable smart devices,

practically half of respondents (44 subjects, 49.4%) indicated price as a main obstacle. Such responses were obtained mainly from people that do not own a device. Another important aspect was the low value of information provided by devices. This response was selected by 19 individuals, that comprises 21% of all respondents. Other aspects, such as data privacy, poor design, and low frequency of use were of lower concern by respondents. The latter aspect could be related to the low need for the device: customers just do not know why they should use wearable devices and how they could benefit from it.

Respondents' evaluation ranged easiness of use as a most important characteristic (average score 4.64 out of 5), reliability is the next by importance (average score 4.45). In turn, the brand of the device was evaluated as factors of lowest importance. Alongside, users put relatively low scores to whether the device is certified. In some extent it could be related to the perception of wearable devices as "gadgets", rather than medical instruments. Multifunctionality and price were rated as factors of intermediate importance.

Concerning potential applications of the wearable devices, the majority indicated monitoring of physical activity data (58 individuals, 65.2%). Information on diet and dietary advice were indicated by 14 respondents (15.7%) while sleep monitoring function – by 7 respondents (7.9%). In a comment section, one respondent indicated that wearable could be integrated within smart home, and another – that it should monitor health status, although did not specify what does it mean. An opposite opinion was that step counting and run distance estimation is enough, and any other function would be redundant. Most of respondents, that selected answer "other" commented, that there is no need in such devices at all.

Among various physiological parameters (except heart rate), respondents were most interested in blood pressure measurements (58 individuals, 65.2%) and body temperature measurements (19 individuals, 21.3%). Respondents expressed low interest in skin humidity measurements (only 4 respondents, 4.5%), and no one respondent chose skin impedance/resistance measurements. This could possibly happen because of the drawback of the questionnaire itself, where relation of skin humidity to stress level, or relation of the body impedance to, say, body fat content, was not emphasized. Another reason is that the public does not consider two latter parameters as important physiological indicators, because such parameters rarely appear in regular medical practice. In the narrative comments, responders mentioned such parameters as heart rate and body fat content.

In general, 77 respondents (86.5%) indicated that they would use wearable medical devices if it could improve their access to information about their health.

Concerning the design of the devices, most of the responders prefer to have the device in the form of wristwatches (66 individuals, 74.2%) or jewelry (14 respondents, 15.7%). Devices integrated into garment or attached to the body were of least interest (4 and 5 respondents only, correspondingly). Most of the responders (74 individuals, 83.1%) agreed that the device should be compatible/connected to the smartphone. Such a result may be related to the present situation on the wearable device market – most devices, advertised by big companies, are smart watches.

Table 1. Responses to the substantive questions.

Question	Response	
What are the obstacles to the purchase of wearable smart devices	price	44 (49%)
	low usefulness of data	19 (21%)
	other	8 (9%)
	low privacy of data	7 (8%)
	low frequency of use	6 (7%)
	poor external design	5 (6%)
What features of wearable devices are more important (average Likert scale estimate)?	ease to use	4.6
	reliability	4.5
	price	4.1
	multifunctionality	4.0
	certification	3.8
	brand	3.3
What information would you like to get from your device?	physical activity data	58 (65%)
	dietary info and advice	14 (16%)
	other	10 (11%)
	sleep monitoring data	7 (8%)
What medical data would you like to take with smart devices?	blood pressure	58 (65%)
	body temperature	19 (21%)
	other	8 (9%)
	skin humidity	4 (5%)
	skin resistance	-
From which source would you like to get information on medical portable devices (average Likert scale estimate)?	medical doctors	4.0
	experienced users	4.0
	manufacturers	4.0
	Internet reviews/feedback	3.3
	state health organizations	3.2
	internet trading platforms	3.2
	sellers, incl. Pharmacies	3.1
	health insurance companies	2.7
Would you use a medical device if it would improve your access to information about your health?	yes	77 (87%)
	no	12 (13%)

(continued)

Table 1. (continued)

Question	Response	
What design of portable medical measuring device would you prefer?	wristwatch	66 (74%)
	jewelry	14 (16%)
	body patch	5 (6%)
	garment	4 (4%)
Does the smart device need to be compatible with / connected to a smartphone?	yes	74 (83%)
	no	15 (17%)

Respondents consider medical doctors and experienced users as the most important information sources about wearable medical devices. This observation is important for wearable device manufacturers planning advertisement campaigns. Direct information from manufacturers has got high credits as well.

In a free comment section, one of the respondents proposed that the device must have a humble design to prevent theft, as well emphasized importance of easy to use and intuitive interface, as a need to search for manual could repulse user. This is especially important if the smart device is used by children or the elderly. The devices must be self-sufficient, i.e., be able to operate without connection to e.g., smartphones, as not everyone who needs a medical wearable can afford or is able to handle a smartphone. Minimization of unnecessary communications with other devices helps maintain data security. Also, such wearable devices could include SOS signal feature. Another responder noted that smart watches still could not be called medical measuring devices because of low measurement accuracy.

4 Discussion and Conclusion

The present research provided a first insight into the expectations of Latvian potential users of wearable medical devices. The main limitation of the study is the limited selection of the respondents in the studied group, which could not be considered as fully representative for the Latvian population. The use of the authors' social media as the primary questionnaire dissemination channel limited the survey to economically active individuals who could afford modern communication devices or computers, were active Internet and social media users, and came from the same social stratum as the authors. This is confirmed by the respondents' age (average age 36.7 years) and occupation (mainly employees) profile. Hereby, seniors were, in fact, excluded despite being the primary potential users of wearable medical technologies. Additionally, the questionnaire did not inquire about the opinions of users who are relatives of seniors. Another limitation is that the study does not involve medical professionals as potential facilitators of the use of the wearable medical devices. These groups should be included in future studies.

Despite such limitations, the study revealed some interesting findings. The results, received from Latvian respondents differs from those, obtained in similar studies. For

instance, responders in Park's study estimated perceived easiness to use as the most important factor, but perceived cost – the least important factor that determines users' intentions to use smart wearable devices [14]. Huarng, Hui-Kuang Yu and Lee, too, indicated, that costs concerns have negligible effect on the decision to use wearable device [15]. In contrast, Latvian respondents considered costs as a main obstacle, preventing them from purchase of wearable device. This difference may simply be due to the formulation of the question: the questions in [14, 15] asked about the factors influencing the decision to purchase the device, whereas in the present research, the question focused on what prevents it. It is important to note that Latvian respondents identified ease of use as the most important feature of the device.

Another important issue revealed by the study was estimation of the user perception of usefulness of the wearable device. The narrative comments, provided by responders, demonstrated that healthy individuals perceive smartwatches or fitness belts, mostly used for fitness activity tracking, as wearable medical devices. Out of 89 responses, only one respondent mentioned Holter monitor, and four – blood glucose meters. That could explain why some respondents do not see any need to use wearable devices unless they are not ill or do not want to track their fitness activities.

The findings of the research helped formulate a list of problems that manufacturers could be advised to address. The most important issues are: 1) non-obtrusiveness of the device, 2) a simple user interface, and 3) the device's ability to operate without a connection to a smartphone. Among the various physiological parameters, monitoring blood pressure would be of high value to users.

Summarizing, one could note, that Latvian respondents generally could use wearable devices if they see clear benefits and find the costs of the devices affordable.

Acknowledgements. This study was funded by the European Regional Development Fund, the Operational Program "Growth and Employment" 1.1.1. Specific support objective "Increase the research and innovation capacity of Latvian scientific institutions and the ability to attract external funding by investing in human resources and infrastructure" 1.1.1.2. Within the measure "Support for post-doctoral research" in the framework of the project No.1.1.1.2/ VIAA/2/18/343 "The methodology for the commercialization of innovative biomedical devices and the evaluation of the productions financing model".

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zheng J, Shen Y, Zhang Z, Wu T, Zhang G, Lu H (2013) Emerging wearable medical devices towards personalized healthcare. In: Proceedings of the 8th International Conference BodyNets '13. ICST, Brussels, pp 427–431
2. Rodriguez-Villegas E, Iranmanesh S, Imtiaz SA (2018) Wearable medical devices: high-level system design considerations and tradeoffs. *IEEE Solid-State Circuits Mag* 10(4):43–52
3. Haghi M, Thurow K, Stoll R (2017) Wearable devices in medical internet of things: scientific research and commercially available devices. *Healthcare Inf Res* 23(1):4–15
4. Olsen LS, Nielsen JM, Simoný C, Kjær TW, Beck M (2021) Wearables in real life: a qualitative study of experiences of people with epilepsy who use home seizure monitoring devices. *Epilepsy Behav* 125:108398

5. Turakhia MP et al (2013) Diagnostic utility of a novel leadless arrhythmia monitoring device. *Am J Cardiol* 112(4):520–524
6. Mück JE, Ünal B, Butt H, Yetisen AK (2019) Market and patent analyses of wearables in medicine. *Trends Biotechnol* 37(6):563–566
7. Isakadze N, Martin SS (2020) How useful is the smartwatch ECG? *Trends Cardiovasc Med* 30(7):442–448
8. Money AG, Barnett J, Kuljis J, Craven MP, Martin JL, Young T (2011) The role of the user within the medical device design and development process: medical device manufacturers' perspectives. *BMC Med Inform Decis Mak* 11:15
9. IJzerman MJ, Koffijberg H, Fenwick E, Krahn M (2017) Emerging Use of Early Health Technology Assessment in Medical Product Development: A Scoping Review of the Literature. *Pharmacoeconomics* 35(7):727–740. <https://doi.org/10.1007/s40273-017-0509-1>
10. Woudstra K, Tummers M, Rovers MM, Reuzel R (2021) Innovators' views on involving users and patients in surgical device development: a qualitative interview study. *BMJ Open* 11(8):e050801
11. Yang F, Al Mahmud A, Wang T (2021) User knowledge factors that hinder the design of new home healthcare devices: investigating thirty-eight devices and their manufacturers. *BMC Med Inform Decis Mak* 21(1):166
12. Ferguson C, Hickman LD, Turkmani S, Breen ,P Gargiulo G, Inglis SC (2021) “Wearables only work on patients that wear them”: Barriers and facilitators TO the adoption of wearable cardiac monitoring technologies. *Cardiovasc Digital Health J* 2(2):137–147
13. Kim YW, Yoon SH, Hwangbo H, Ji YG (2017) Development of a user experience evaluation framework for wearable devices. In: Zhou J, Salvendy G (eds) ITAP 2017, vol 10298. LNCS. Springer, Cham, pp 53–67. https://doi.org/10.1007/978-3-319-58536-9_5
14. Park E (2020) User acceptance of smart wearable devices: an expectation-confirmation model approach. *Telematics Inform* 47:101318
15. Huarng KH, Yu THK, Lee CF (2022) Adoption model of healthcare wearable devices. *Technol Forecast Social Change* 174:121286
16. Official Statistics Portal. <https://stat.gov.lv/en/statistics-themes/population/population/247-population-and-population-change>. Accessed 25 Feb 2021
17. Sample Size Calculator by Raosoft, Inc. <http://www.raosoft.com/samplesize.html>. Accessed 25 Feb 2021
18. VisiDati. Create your online survey. <http://en.visidati.lv/>. Accessed 09 Sept 2021
19. Caron A, Vandewalle V, Marcilly R, Rochat J, Dervaux B (2021) The optimal sample size for usability testing, from the manufacturer's perspective: a value-of-information approach. *Value Health* 25(1):116–124
20. Heponiemi T, Jormanainen V, Leemann L, Manderbacka K, Aalto AM, Hyppönen H (2020) Digital divide in perceived benefits of online health care and social welfare services: national cross-sectional survey study. *J Med Internet Res* 22(7):e17616
21. Di Giacomo D, Ranieri J, D'Amico M, Guerra F, Passafiume D (2019) Psychological barriers to digital living in older adults: computer anxiety as predictive mechanism for technophobia. *Behav Sci* 9(9):96



Subjective Evaluation of Phantoms in X-ray Breast Imaging Research: Insights from Bulgaria

Tsvetelina Teneva¹ , Antonio Sarno² , Nikolay Dukov¹ ,
Vasileios Elefteriadis³ , Zhivko Bliznakov¹ , Panagiotis Papadimitroulas³ ,
and Kristina Bliznakova¹

¹ Medical University of Varna, Varna, Bulgaria
{zhivko.bliznakov, kristina.bliznakova}@mu-varna.bg

² University of Milan, Milan, Italy

³ BIOEMTECH, Athens, Greece

Abstract. This work aims at defining a validation approach to subjectively assess the realism of breast phantom images acquired with x-ray systems. Various breast phantoms were used to address specific questions. First, the realism of a breast neoplasm model was evaluated. Second, different materials for phantom production were evaluated for x-ray contrast, sharpness, inhomogeneity, and artifacts. Third, 3D printing techniques for manufacturing breast tissues were explored. Lastly, computational models were assessed by comparing manual and automatic segmentations of the breast tissues. Two distinct grading scales were introduced to aid radiologists in evaluating the phantoms. The first scale assesses computational phantoms based on the realism of glandular tissue distribution, morphology, and the preservation of fine parenchymal structures. The second scale is designed to support the validation of physical phantoms, focusing on the realism of the generated x-ray image. The scales have been applied to the selected cases and will be further used in studies involving the production of anthropomorphic phantoms.

Keywords: radiologist evaluation · anthropomorphic phantoms · breast imaging · breast neoplasm · segmentation

1 Introduction

Development, testing and validation of novel breast imaging techniques rely on the use of anthropomorphic breast phantoms [1, 2]. In order to be as realistic as possible the validation of the images by experienced radiologist is crucial, in particular as regard to the validation of the realism of the produced images, with evaluation almost exclusively being carried out using objective quantitative methods [3].

Medical University of Varna leads a Horizon 2020 project **PHENOMENO** (<https://phenomeno.eu/>) aiming at the development of a methodology for creation of computational anthropomorphic breast phantoms based on Magnetic Resonance Imaging (MRI). In the project framework, the role of the radiologist is to segment manually the different

breast tissues in order to produce both computational and physical breast phantoms and to provide feedback for the machine-learning algorithm on the location and the type of the pathologic finding inside the breast. Once the breast phantoms are created, the radiologist is responsible for their subjective validation. This is a challenging task if taking into account that both computational and physical phantoms may not consistently align with anatomical and radiological accuracy.

Adequate assessment of non-clinical, computational x-ray and MR images from experienced radiologists is still not possible. Radiologists in the field of in-silico medicine are still troubled in interpretation of images that are computer modelled or images from phantoms. They should become familiar with different contrast, sharpness and non-anatomical structure of the model scanned. In this context, the key questions posed are: How can the 2D and 3D images obtained from anthropomorphic breast phantoms be effectively interpreted, and how can the realism of the simulated structures be accurately evaluated?

The objective of this work is to establish paradigms for subjective validation of developed physical breast phantoms. To achieve this, we introduce four cases of in-house developed breast phantoms. Crucial step in our process involves validation by the team radiologist before and after x-ray scanning.

2 Materials and Methods

We selected four distinct phantoms and image data set designed and developed by our research team [4–8]. Each case represents a particular aspect of the validation process, spanning from the generation of the computational model to the manufacturing of the physical phantom. A specific validation research question is related to each case.

3 Case 1: Neoplasm Representation

Research Question: Which Model Gives the Most Realistic Representation of a Neoplasm?

Phantom Description: Comparison between models and real breast cancer detected on mammography images.

Two physical breast phantoms were created using two 3D printing technologies: fused-deposition modelling (FDM) and stereolithography (STL) (Fig. 1a, b). The different parts of the breast were computationally designed by using the in-house developed software *BreastSimulator* [5], which were subsequently transformed into STL files. The different parts of the phantoms – skin, gland, adipose and lesions were printed separately and then assembled in phantoms.

Mammographic images of the phantoms are obtained at a digital mammography system, by using Automatic Exposure Control (AEC) at 34 kV tube voltage with an Rh/Ag anode-filter combination. Two scanning scenarios for the phantoms were implemented. In the first scenario, the empty spaces between the phantom's components were left unfilled (Fig. 2a), while in the second scenario, the phantoms were filled with water (Fig. 2b).

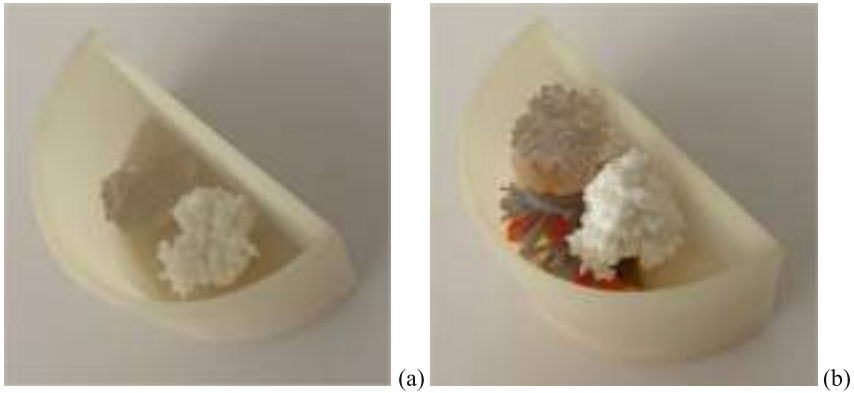


Fig. 1. Breast phantoms comprising: (a) an external envelope and two printed lesions; (b) with models of gland and adipose tissue to reproduce the anatomical noise in x-ray images.

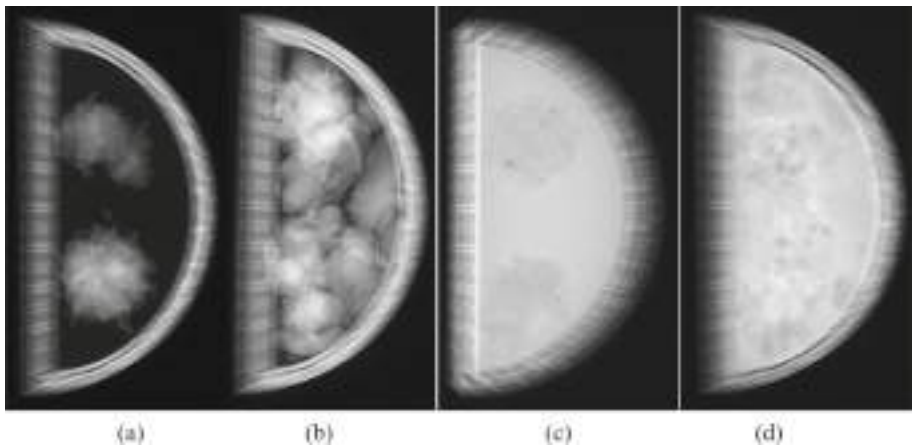


Fig. 2. Mammography images of the phantoms in Fig. 1(a, b) containers filled with air/no filling; and (c, d) containers, filled with water.

4 Case 2: Material Selection

Research Question: Which Material is Suitable for Breast Lesion Representation?

Phantom Description: Several lesion models were manufactured, selected from the Maxima database [6]. A screenshot of a lesion model is shown in Fig. 3a. This computational model was printed from two different materials and placed in a container (Fig. 3b) made of two components: inner white container, printed from polylactic acid (PLA) simulating a compressed breast and an outer black component printed from PLA, simulating breast skin [7].

Images of the phantoms were obtained on a digital mammography system in AEC mode [7]. The applied anode voltage was set to 28 kV, with a tungsten anode and a rhodium filter configuration. Different phantoms were developed by changing the used

materials. As regard to the filling material, we used either water or paraffin (the filling material refers to the space surrounding the tumor model within the container) and different solid materials for cancer (different types of 3D printing materials). Figure 3c shows two mammographic images of two phantoms, each featuring tumors fabricated from different materials.

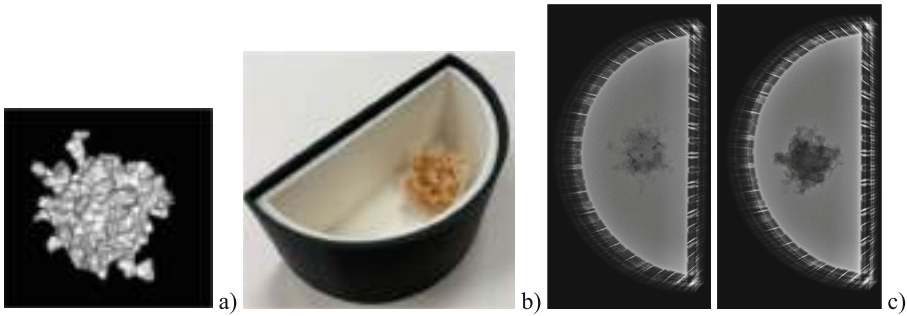


Fig. 3. Phantom with an irregular breast lesion: a) computational lesion model selected from the Maxima database [6]; b) 3D-printed breast phantom comprising a dual-layered envelope and a physical lesion replica. c) mammography images of the phantom filled with paraffin and two different materials used for tumor.

5 Case 3: Technique Selection

Research Question: Are Clinical Images and Phantom's X-Ray Images Comparable?

Phantom Description: 3D physical breast phantoms produced via FDM and inkjet printing techniques are reported in Fig. 4. They were manufactured from breast MR images [8]. The FDM breast phantom is based on a new printing technology based on variable extrusion rate of the plastic material on the basis of the desired local density and by using one single material (PLA in this case) [4]. In the second case, the phantom was printed on a paper sheet via inkjet printer, and different concentration of the ink permitted to represent the different tissues within the organs. X-ray images of the phantoms were acquired via computed tomography (CT) Siemens Somatom Force and abdomen standard protocol at 70 kV and 130 mAs. Reconstructed slices were 0.5 mm thick and presented corona pixel pitch of 0.45 mm.

6 Case 4: Computational Model

Research Question: Is it Possible to Achieve an Outstanding Segmentation of the Breast Glandular Tissue and of the Neoplasm?

Case Description: MR images from a patient with histologically proven breast cancer was selected. After contrast medium application breast cancer is easily detectable on T1 weighted sequence as the brightest structure. Once the cancerous tissue is detected,



Fig. 4. 3D technique selection: (a) paper-based (left) and FDM-based (right) physical breast phantoms manufactured from patient MR breast images. (b) (A) MRI of the patient's breast for diagnostic purpose (B) the computational model of the breast phantoms. CT slices of (C) the FDM 3D printed phantom and (D) the paper phantom.

its segmentation is easily achievable. The cancer segmentation may be performed either automatically or manually with prevalence on the manual segmentation (Fig. 5). In our study, manual segmentation was performed by the radiologist using a Wacom Pen Tablet and the mDicom application (MicroDicom Ltd, Bulgaria). Automatic segmentation was carried out using a Deep Neural Network. The design and implementation utilized MONAI, an open-source PyTorch-based framework [9].

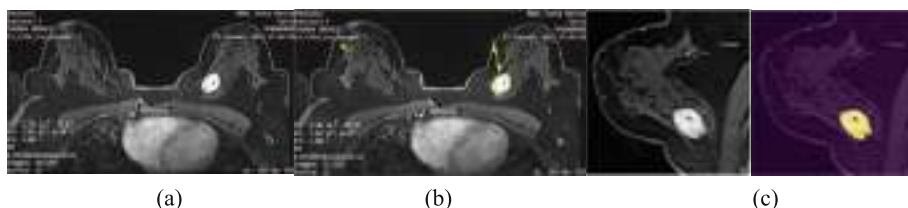


Fig. 5. a) Clinical breast MRI- T1 weighted sequence, contrast enhanced, axial view, b) manual segmentation of tumor only, (c) clinical breast MRI- T1 weighted sequence, contrast enhanced, sagittal view, and predicted automatic segmentation of tumor only.

7 Results and Discussion

7.1 Subjective Evaluation of Computational Phantoms

The subjective analysis and validation of the generated computational phantoms require the evaluation of the presence of typical structure and of the morphology of the breast parenchyma. In particular, the segmentation process may compromise the anatomical structures that are visible in the initial clinical images used for the model generation; on the other hand, classical approaches in image segmentation do not permit the introduction of structures not visible in the initial clinical image. For this reason, we propose a similarity scale meant to compare the generated computational breast phantoms to the

initial clinical image. The proposed scale is reported in Table 1, where structures in generated models are compared to real MR images. The scale evaluates two key aspects: (1) the distribution and morphology of the segmented glandular tissue, and (2) the degree of preservation of fine details such as ligaments, interstitium, and ducts. Each aspect is assessed using a similarity grading system, ranging from 1 (very low correspondence) to 5 (full correspondence). Based on this scale, images in cases 1 2, and 3 are evaluated in Table 4 (Table 2).

Table 1. Similarity scale for radiological evaluation of generated breast models.

Grade	Glandular tissue evaluation	Fine details preservation
1	The distribution of the glandular tissue and morphology in the segmented image do not correspond to clinical MR image	Ligaments, interstitium, ducts and small details visible in the clinical MR images are not visible at all or mostly not preserved on the segmented images
2	The distribution of the glandular tissue and morphology in the segmented images largely differ from those in MR images	Ligaments, ducts and small details are lost, fragmented or largely modified in the segmented images
3	The distribution of the glandular tissue partly corresponds to glands in clinical MR images. In some portions of the segmented image, the morphology cannot be recreated as the morphology of the breast	Larger part of ligaments, ducts and small details are still present in the segmented images, but their morphology is mainly altered (they are either larger or smaller than in MR images or fragmented)
4	The distribution of the glandular tissue reflects the glands on clinical MR images, but its morphology differs from those represented in MR images	Most of ligaments, ducts and small details are present in the segmented images but with altered morphology
5	The distribution and morphology of the segmented glandular tissue truly recreates those in clinical MR images	All ligaments, ducts and small details visible in clinical MR images are still present after image segmentation with unaltered morphology

Table 2. Exemplary radiological assessment of generated breast models.

Case	Grade	Comments from the radiologist
2	Tumor related image characteristics Grade 4; Details 1	Evaluation of internal tumor characteristics is made based on my experience with previous tumor models
4	not applicable	Soft tissue segmentation only

7.2 Subjective Evaluation of Different Aspects from Physical Phantoms

As regard to the subjective evaluation of physical phantoms, researcher must be able to answer the questions we have asked for cases 1, 2, 3 namely: 1) is the physical model of the neoplasm realistic; 2) which material is the most suitable for breast tissue manufacturing; 3) is it possible to compare clinical images and phantom's x-ray images. In order to dive into these three aspects, we must evaluate the realism of the obtained images acquired with clinical scans. For clinical breast images from mammography, MR, CT, the radiologist is able to distinguish the following anatomical structures: 1) connective tissue consisting of mammary glands and ducts, small vessels, lymph vessels and nerves, 2) adipose tissue consisting of fatty structures, 3) greater vessels (arteries and/or veins) in the periphery of the breast and 4) skin (again soft tissue component). This evaluation relies on subjective observations guided by structured criteria, including size, margins, margin visibility, contrast ratio, target-to-background sharpness, and the presence of artifacts within the target. Additionally, the perceived realism of anatomical and pathological structures, such as neoplasia, are also assessed. The proposed questionnaire for radiological evaluation of breast phantom is included in Table 3.

Table 3. Questionnaire for radiological evaluation of breast phantoms.

Question	Answers
1. Does the image contain any visible lesions?	Yes No Uncertain
2. Does the lesion appear realistic?	Yes, completely realistic Partially realistic Not realistic
3. What is the type of the lesion observed?	Benign, Malignant, Calcification, Cystic, Other
4. Can you clearly differentiate between the various breast components?	Yes: skin, adipose, glands are clearly distinguishable Yes: skin, adipose, glands, neoplasm are clearly distinguishable Partially yes: skin and adipose are clearly distinguishable No, they are indistinguishable
5. Are the modeled breast components realistic in appearance?	Yes, highly realistic Somewhat realistic Not realistic at all
6. Does the breast model, as a whole, appear realistic?	Yes, very realistic Moderately realistic Not realistic

Using the proposed questionnaire, the radiologist evaluated the selected cases as follows: Case 1 involved the phantom shown in Fig. 1b and the image in Fig. 2d; Case 2; and Case 3, which featured the FDM breast phantom labeled as 'C' in Fig. 4b.

The evaluations performed by several experts can be combined through weighted quantitative scoring, as the most suitable phantom obtains the highest score. A number

Table 4. Exemplary radiological assessment of breast phantoms.

Question	Case 1	Case 2	Case 3
1.Does the image contain any visible lesions?	No	Yes	No
2.Does the lesion appear realistic?	Partially realistic	Partially realistic	No lesion
3.What is the type of the lesion observed?	Solid	Non-solid	Not visible
4.Can you clearly differentiate between the various breast components?	Yes: skin, adipose, glands are clearly distinguishable	Partially yes: skin and adipose are clearly distinguishable	Yes: skin, adipose, glands are clearly distinguishable
5.Are the modeled breast components realistic in appearance?	Somewhat realistic	Somewhat realistic	Not realistic
6.Does the breast model, as a whole, appear realistic?	Moderately realistic	Not realistic	Moderately realistic

of five radiologists is sufficient to consider the evaluation results representative for the grading of a specific phantom.

8 Conclusions

In-silico studies and studies with radiological phantoms hold significant potential for testing and facilitating the approval of innovative apparatuses for breast cancer diagnosis, offering reduction in cost, time, and patient exposure risks. Radiologists play a crucial role in advancing breast imaging research, emphasizing the need for well-designed subjective evaluation methods. This study has developed frameworks for the subjective evaluation of newly created and manufactured breast phantoms, which will be applied in the ongoing assessment of new anthropomorphic models. These efforts accelerate advancements in the field of radiology, which is made possible through the collaboration of an interdisciplinary team. The close partnership between engineers, researchers, and radiologists is a distinctive strength in Bulgaria, offering substantial benefits to the field.

Acknowledgement. This work is supported by PHENOMENO action. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101008020.

The authors have no competing interests to declare.

Ethics Approval. This study was approved by the Ethics Committee of Medical University of Varna (Approval number 102/22.04.2021).

References

1. Bliznakova K (2020) The advent of anthropomorphic three-dimensional breast phantoms for X-ray imaging. *Phys Med* 79:145–161
2. Sarno A, Valero C, Tucciariello R, Dukov N, Costa P, Tomal A (2023) Physical and digital phantoms for 2D and 3D x-ray breast imaging: review on the state-of-the-art and future prospects. *Rad Phys Chem* 204:110715
3. Wang J et al (2023) Image simulation and realism evaluation for mammography and tomosynthesis based on the detailed breast phantom. *Rad Phys Chem* 213:111241
4. Dukov N, Bliznakova K, Okkalidis N, Teneva T, Encheva E, Bliznakov Z (2022) 2022 Thermo-plastic 3D printing technology using a single filament for producing realistic patient-derived breast models. *Phys Med Biol* 67(4):045008
5. Bliznakova K, Sechopoulos I, Buliev I, Pallikarakis N (2012) BreastSimulator: a software platform for breast x-ray imaging research. *J Biomed Graph Comput* 2:1–14
6. Bliznakova K et al (2019) Development of breast lesions models database. *Phys Med* 64:293–303
7. Toschkova O (2023) Anthropomorphic model of a breast with lesions for X-ray studies. In: Department of Medical Equipment, Electronic and Information Technologies in Healthcare, (Varna, Bulgaria: Medical University of Varna), diploma thesis
8. Bliznakova K et al (2024) A comparison study of two low-cost 3D printing technologies for manufacturing radiological breast phantoms from MRI breast images. *IJRR* 22(4):883–890
9. Cardoso MJ, et al (2022) MONAI: An open-source framework for deep learning in healthcare. [arXiv:2211.02701](https://arxiv.org/abs/2211.02701)



EGGS: Efficient Gathering and Structuring of Avian Egg Datasets

Lukasz Roszkowiak¹(✉), Pawel Pstrokowski², Wojciech Wojcik²,
Krzysztof Damaziak², and Anna Korzynska¹

¹ Nalecz Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences, Warsaw, Poland

lroszkowiak@ibib.waw.pl

² Warsaw University of Life Sciences, Warsaw, Poland

Abstract. This paper presents EGGS, a novel and purpose-driven image processing approach designed specifically for creating a standardized and comprehensive dataset of avian egg images. The proposed approach integrates advanced techniques for image preprocessing, orientation detection, and the removal of damaged specimens, addressing key challenges in oological data analysis. With a primary focus on enhancing the quality of avian egg datasets, this study highlights the potential for combining advanced image processing techniques with deep learning models to address challenges in ecological data analysis. Beyond basic preprocessing, EGGS incorporates a robust orientation detection component and precise contour analysis, offering a focused and efficient framework for egg annotation. This innovative solution is tailored to support targeted ecological and heritage research, providing a unique tool for advancing ornithological studies. The resulting dataset not only enables downstream analyses but also serves as a critical resource for ornithological research, opening pathways for deeper insights into avian reproductive strategies, evolutionary adaptations, and environmental resilience.

1 Introduction

1.1 Background

Understanding egg biology, or oology, is key to exploring the ecological and evolutionary strategies of egg-laying species [1]. Eggs are a remarkable adaptation, serving as protective, nutritive vessels for developing embryos. Their diverse shapes, colors, and patterns reflect responses to environmental pressures, predation, and reproductive strategies. As a cornerstone of ornithology, oology provides critical insights into biodiversity, adaptation, and the interactions between organisms and their environments, bridging evolutionary biology and conservation science [2].

The creation of a standardized dataset of avian egg images plays a pivotal role in advancing research across diverse scientific domains. Such datasets not

L. Roszkowiak and P. Pstrokowski—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
P. Ladyzynski et al. (Eds.): NBC/PCBBE 2025, IFMBE Proceedings 131, pp. 320–328, 2025.
https://doi.org/10.1007/978-3-031-96538-8_28

only serve as a repository for visual information but also pave the way for leveraging machine learning applications, ecological studies, and conservation efforts. Robust datasets enable researchers to discern patterns, variations, and trends in avian egg characteristics, contributing to a deeper understanding of avian biodiversity and evolutionary dynamics [3]. Hence, a comprehensive dataset that captures the diverse array of avian eggs becomes an invaluable resource for researchers.

In this context, the development of advanced image processing techniques becomes imperative. Through the proposed image processing approach, we seek to address the challenges posed by the labor-intensive and time-consuming tasks of manual labeling, segmentation, and transformation of images. The ability to preprocess images effectively and determine the orientation of the eggs, not only enhances the quality of datasets but also empowers researchers with standardized and aligned visual representations. This procedural advancement forms the backbone of our research, aiming to contribute to the broader field of avian biology and ornithology.

1.2 Objectives

The overarching objectives of this study are to:

- Develop an advanced image preprocessing algorithm tailored for avian egg dataset creation.

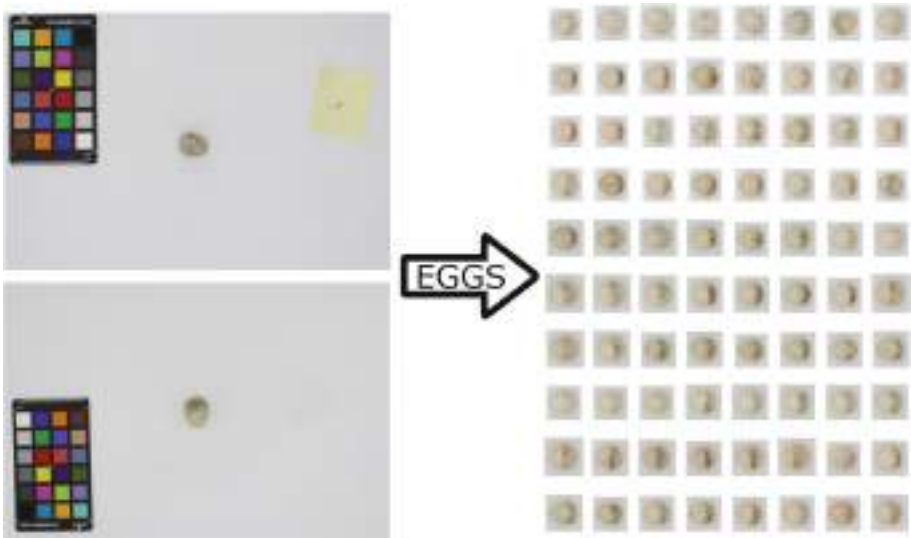


Fig. 1. Overall image file processing with EGGS. An example of a full-scale image containing the object of interest along with additional elements (on the left); the result of processing using the EGGS method, where the object of interest is extracted, the image is cropped, and the object's orientation is corrected (on the right).

- Resulting dataset should be ready for use with machine learning methods.
- Accurately detect and categorize the orientation of eggs within the dataset.
- Detect and remove cracked eggs from analysis.
- Present a step-by-step breakdown of the method and its components.

2 Methods

The study began with the development of a general approach for automatically generating an image database from collected photographs (Fig. 1). Such images, commonly captured following standardized protocols in oological studies, often include additional objects and information, such as a colorimeter, ruler, or handwritten notes, alongside the primary subject. To prepare the images for classical or AI-based analysis, preprocessing is essential to isolate the object of interest.

The initial step in the preprocessing pipeline involved detecting and extracting the primary object from the image. An automated method was proposed for this task, ensuring efficient segmentation of the target from surrounding elements. Once the object was identified, its shape and position were verified to ensure accurate extraction of the specific image fragment containing only the egg.

2.1 Image Preprocessing

The preprocessing algorithm commences with the conversion of RGB images into the LAB color space [4]. Subsequently, the L (Luminescence) channel undergoes thresholding using the Triangle method [5], resulting in a binary image highlighting potential egg regions. To mitigate irregularities and holes, a flood filling technique is employed, followed by morphological operations such as opening and closing for further refinement.

2.2 Contour Analysis

Potential egg shapes are identified by analyzing the image to detect contours, focusing on the outermost edges and simplifying their outlines to capture the essential shape of each egg. This approach ensures precise identification while minimizing the influence of noise and artifacts present in the image. The algorithm meticulously processes the identified contours, evaluating criteria such as area, aspect ratio, and circularity to filter out non-egg objects. This step concludes by producing binary mask of object of interest.

2.3 Orientation Detection

Regardless of assumed protocol for image acquisition there are always examples that would divert from expectations. It happens that the egg is flipped or simply slightly rotated. Oftentimes it is impossible to set it on placeholder ideally due to its ingenious shape or deformation.

For each potential egg, the algorithm assesses its position within the image and calculates aspect ratio and circularity metrics. Egg objects meeting predefined thresholds for aspect ratio and circularity are classified as either horizontal or vertical based on their orientation. An ‘other’ category is introduced for eggs that do not meet these criteria, and they could be stored separately or included in the final dataset, enhancing the dataset’s inclusivity. The detailed orientation detection further refines the dataset, preparing it for use in training machine learning models that can distinguish between different egg orientations. By standardizing egg orientation, we avoid scenarios where machine learning models might incorrectly use orientation as a classification criterion instead of focusing on relevant features.

It was assumed that to minimize potential distortions introduced by interpolation during image preprocessing, eggs positioned at an angle of less than 10° relative to the horizontal axis are left unrotated. This threshold ensures that minimal angular adjustments are applied to samples that are already near-horizontal, thereby preserving their original geometric integrity.

For eggs identified as being predominantly horizontal in orientation, rotation of 180° is applied only when necessary to standardize their positioning. Additionally, these eggs are rotated as required to ensure that their sharp end is consistently oriented toward the left-hand side of the image (Fig. 2). This standardized alignment facilitates uniformity across the dataset, which is critical for applications requiring explainable AI (XAI) [6]. By enforcing a consistent spatial orientation, the interpretability of model decisions is enhanced, allowing for more meaningful and reliable explanations of model behavior.

Eggs classified as vertically oriented undergo a two-step adjustment process. Initially, they are rotated by 90° to achieve a horizontal orientation. Rotating by 90° is less harmful to image quality compared to other angles, as it preserves the pixel grid structure without introducing interpolation artifacts, thereby maintaining the visual integrity of the image. Following this, further fine-tuning of their alignment is conducted, if needed, to ensure conformity with the standardized positional criteria.

While standardized alignment is essential for XAI applications, a different approach could be used during model training. Flipping and rotation could be applied as part of standard image augmentation procedures to enhance the diversity of the training data. These augmentations help the model generalize better by exposing it to variations in egg orientation and position, which can occur in real-world scenarios. This combination of augmentation during training and enforced standardization for XAI ensures both robust model performance and enhanced interpretability of its outputs [7].

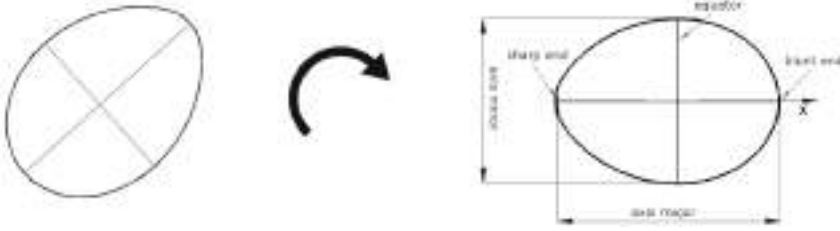


Fig. 2. Egg image orientation processing.

2.4 Crack Detection

Cracked and broken eggshells (examples shown in Fig. 3) can significantly interfere with the accurate classification and feature extraction of samples in image-based analyses. The presence of such anomalies introduces noise into the dataset, which can adversely affect the performance and reliability of downstream processes, including machine learning models trained on the data. Consequently, the exclusion of these undesired samples is critical to ensure the integrity and accuracy of the dataset.

To address this issue, an AI-based approach utilizing the EfficientNet [8] architecture was implemented to identify and filter out images containing cracked or broken eggshells. EfficientNet was chosen due to its state-of-the-art performance in image recognition tasks, offering a balance between computational efficiency and accuracy.

To leverage pre-existing knowledge, the model was initialized with pretrained weights (ImageNet), excluding the top classification layers. These fully connected layers were custom-designed and adapted to fit the specific requirements of our binary classification task, which focused on distinguishing damaged (cracked) from undamaged eggshells. Initially, the base layers of the model were frozen to retain the pretrained feature extraction capabilities and prevent overfitting during the initial training phase.

The model was initially trained on a publicly available dataset of chicken eggs [9], which provided a general representation of intact and damaged eggshells. The conglomerate dataset (containing multiple species) used for training consisted of 888 images, which were divided into training, validation, and test subsets following standard data partitioning practices. The model was trained for 30 epochs using the Adam optimizer and categorical cross-entropy as the loss function, which is well-suited for tasks involving multi-class or binary classification.

Following this stage, the model was unfrozen, allowing all layers to be trainable, and fine-tuned with a reduced learning rate for an additional 10 epochs. This two-step training approach helped refine the model's performance, ensuring that the pretrained features were optimally adapted to the nuances of our specific dataset.



Fig. 3. Exmple of cracked eggs that should be excluded from dataset.

This approach ensures that the dataset is composed exclusively of high-quality samples, thereby optimizing the conditions for subsequent analyses and improving the overall performance of the classification pipeline.

3 Results

3.1 Orientation

The algorithm’s effectiveness is demonstrated through the preprocessing and orientation detection of diverse avian egg images, showcasing its ability to standardize orientation and enhance overall image usability. The resulting dataset provides a valuable resource for researchers, offering a comprehensive collection of avian egg images suitable for various analyses.

The developed method for automatic extraction of objects of interest achieved a success rate of more than 85% on our dataset [10]. By applying this automatic process, an image dataset containing more than 3,000 examples was finally produced, ready to be used in the neural network learning process. The resulting dataset contains examples of eggs of 5 different species of birds of the order Lanius [10] with the abundance shown in Fig. 1.

To further demonstrate the generalization ability of the proposed method, we conducted tests on two other publicly available datasets featuring images of eggs acquired in a scientific setting. From the *Egg Crack Detection* dataset available on Roboflow [11], we randomly selected 100 images. The method successfully cropped all images; however, it failed to correct the orientation in 9 out of 100 samples.

Additionally, we tested the method on images of King Rail and Common Moorhen eggs from a study on interclutch variability [12, 13]. While these images were preprocessed to feature a homogeneous black background, their initial orientation differed from our intended standard and required correction (Table 1).

Table 1. Performance evaluation of the proposed method: accuracy results across different datasets.

	dataset 1[10]	dataset 2[11]	dataset 3[13]
accuracy	85%	89%	99%
source	our own	Roboflow	Dryad

3.2 Crack Detection

The EfficientNet model was employed as the base architecture, chosen for its state-of-the-art performance in image classification tasks. The final model demonstrated exceptional performance, achieving a 97% accuracy on the test set. To further validate its effectiveness, the model's performance was evaluated on a separate subset consisting exclusively of cracked images from the Lanius dataset. Out of 34 images in this subset, only a single image was misclassified as undamaged when it was, in fact, cracked.

This rigorous training and evaluation process highlights the model's potential for application in automated quality control requiring high precision.

4 Discussion

The presented findings demonstrate the robustness and utility of the proposed approach for both image preprocessing and crack detection in avian egg datasets. By optimizing image analysis workflows, these advancements create opportunities for broader use in ecological and evolutionary research.

The algorithm for orientation detection and preprocessing proved highly effective, enabling the generation of a standardized dataset of avian egg images. By achieving an orientation success rate exceeding 85%, this method ensures that the dataset is both consistent and of high quality. Such standardization is crucial for downstream applications, as it minimizes variability and enhances the reliability of feature extraction and model predictions.

Furthermore, the automated extraction process significantly reduced the manual effort required, producing a dataset. By automating tasks that were traditionally labor-intensive, such as object identification and segmentation, the process enhanced efficiency and minimized the potential for human error. As a result, a high-quality dataset can be generated in mere minutes, ready for immediate use in neural network training and other analytical applications.

Moreover, the EfficientNet-based crack detection model demonstrated remarkable accuracy (97%) on the test set, highlighting its effectiveness for binary classification of eggshell damage. The transfer learning approach, combined with fine-tuning on a domain-specific dataset, ensured that the model effectively adapted to the nuances of avian egg imagery.

4.1 Limitations and Future Work

While the algorithm excels in standard egg image scenarios, challenges may arise in cases of extreme lighting variations or non-standard egg shapes. Future work could focus on refining the algorithm to address these limitations and expanding its applicability to a broader range of avian species, thus contributing to the continuous evolution of avian egg image datasets.

Low contrast between eggshells and their backgrounds represents another potential limitation, particularly in non-scientific settings with diverse and cluttered backgrounds where the current method may perform suboptimally. To

overcome this, further enhancements to the method or alternative approaches should be explored. For instance, novel object segmentation techniques, such as those based on Convolutional Neural Networks (CNNs), could be employed once sufficient labeled data becomes available.

Additionally, the method was tested on a limited number of species, which raises concerns about its generalizability to eggs with significant variations in size, shape, or coloration. Expanding the scope of the approach to accommodate these variations will be essential for its broader adoption and effectiveness in diverse ecological and conservation applications.

5 Conclusion

This study introduces a novel image processing approach designed for the creation of a comprehensive avian egg images dataset. The results presented in this study highlight the potential for combining advanced image processing techniques with deep learning models to address challenges in oological data analysis. By automating tasks such as orientation standardization and damage detection, the proposed procedure reduces the time and effort required for manual intervention, enabling researchers to focus on higher-level analyses. Beyond serving as a mere collection of images, the dataset represents a valuable resource for exploring avian reproductive strategies and environmental adaptations. Overall, the proposed methods and their demonstrated success represent a significant step forward in leveraging computational tools for ecological and biological research.

References

1. Stoddard, M.C., et al.: Camouflage and clutch survival in plovers and terns. *Sci. Rep.* **6**(1) (2016). <https://doi.org/10.1038/srep32059>
2. McCullough, J.M., Hruska, J.P., Oliveros, C.H., Moyle, R.G., Andersen, M.J.: Ultraconserved elements support the elevation of a new avian family, Eurocephalidae, the white-crowned shrikes. *Ornithology* **140**(3) (2023). <https://doi.org/10.1093/ornithology/ukad025>
3. Stoddard, M.C., Kilner, R.M., Town, C.: Pattern recognition algorithm reveals how birds evolve individual egg pattern signatures. *Nat. Commun.* **5**(1) (2014). <https://doi.org/10.1038/ncomms5117>
4. Zeileis A, Hornik K, Murrell P (2009) Escaping rgbland: selecting colors for statistical graphics. *Comput. Stat. Data Anal.* **53**(9):3259–3270. <https://doi.org/10.1016/j.csda.2008.11.033>
5. Zack GW, Rogers WE, Latt SA (1977) Automatic measurement of sister chromatid exchange frequency. *J. Histochem. Cytochem.* **25**(7):741–753. <https://doi.org/10.1177/25.7.70454>
6. Ali, S., et al.: Explainable artificial intelligence (xai): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* **99**, 101,805 (2023). <https://doi.org/10.1016/j.inffus.2023.101805>
7. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). <https://doi.org/10.48550/ARXIV.1705.07874>

8. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. arXiv (2019). <https://doi.org/10.48550/ARXIV.1905.11946>
9. Botta, B., datta, a.k.: Dataset for real-time crack detection on chicken eggs (2022). <https://doi.org/10.6084/M9.FIGSHARE.21568425.V1>
10. Pstrokoński, P., et al.: Can explainable AI classify shrike (Laniidae) eggs by uncovering species-wide pigmentation patterns? PLoS One **20**(5), e0321532 (2025). <https://doi.org/10.1371/journal.pone.0321532>
11. Cracked egg dataset (2023). <https://universe.roboflow.com/crack-egg-detection/egg-crack-detection-yibfi>
12. Johnson, E.W., McRae, S.B.: Interclutch variability in egg characteristics in two species of rail: Is maternal identity encoded in eggshell patterns? PLOS ONE **17**(1), e0261868 (2022). <https://doi.org/10.1371/journal.pone.0261868>
13. Rail egg dataset (2021). <https://doi.org/10.5061/dryad.6q573n60j>



Digital Measurement Modules for Non-invasive Monitoring of Physiological Parameters Under Extreme Conditions

Ewelina Sobotnicka¹ , Jan Mocha^{1,2} , Grzegorz J. Nowak^{1,3} ,
Grzegorz Badura¹, Aleksander Sobotnicki¹ , and Łukasz Dziuda⁴ 

- ¹ Lukasiewicz Research Network – Krakow Institute of Technology, Zakopianska St. 73,
30-418 Krakow, Poland
ewelina.sobotnicka@kit.lukasiewicz.gov.pl
- ² Department of Digital Systems, Silesian University of Technology, Akademicka St. 16,
44-100 Gliwice, Poland
- ³ Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta St. 40,
41-800 Zabrze, Poland
- ⁴ Military Institute of Aviation Medicine, 54/56 Z. Krasińskiego St., 01-755 Warsaw, Poland

Abstract. The advancement of real time physiological monitoring technologies plays a vital role in ensuring the safety of individuals operating in extreme environments, including pilots, athletes, and professionals exposed to harsh environmental conditions. This article introduces a concept for digital measurement modules designed for non-invasive and continuous monitoring of critical physiological parameters. These include cardiovascular, respiratory, and musculoskeletal system activity. The proposed solutions integrate advanced measurement technologies such as electrocardiography (ECG), impedance cardiography (ICG), electromyography (EMG), and Doppler ultrasonography to assess blood flow. The modules are optimized for mobile applications, offering high precision and resistance to external interferences. Their implementation enables the real-time collection and analysis of physiological data, facilitating early detection of potentially hazardous conditions. This is particularly relevant in scenarios where exposure to extreme gravitational forces, ischemic hypoxia, or other environmental stressors can significantly impact human performance and safety. A key advantage of these modules is the ability to provide continuous and reliable physiological assessment without restricting movement, making them suitable for use in dynamic environments. The collected data can also contribute to research on human adaptation to extreme conditions, helping to refine safety protocols and enhance overall performance in aviation, professional sports, and other high risk industries. The developed technology represents a step forward in physiological monitoring, combining precision, mobility, and real time adaptability, thus offering a valuable tool for both safety and performance optimization.

Keywords: measurement modules · medical devices · physiological parameters · non-invasive monitoring · extreme conditions

1 Introduction

Modern technologies for real time monitoring of physiological parameters play a crucial role in ensuring the safety and health of individuals working in extreme conditions, such as pilots, athletes, or workers exposed to hazardous environmental conditions. Special attention is given to monitoring the cardiovascular, respiratory, and musculoskeletal systems, which are particularly sensitive to changes caused by physical and mental stress in extreme environments. In the context of aviation, changes in the functioning of these systems due to gravitational forces, fluctuating atmospheric pressure, and psychophysical stress can lead to serious health consequences, including fluid balance disorders, decreased blood oxygen saturation, increased blood pressure, and elevated heart rate [1]. To effectively monitor the health of individuals performing tasks in such conditions, it has become essential to develop modern, mobile, and precise measurement systems that enable non-invasive, continuous data collection on their physiological state. In this context, one of the main challenges is to design a solution that ensures high measurement accuracy while remaining resistant to external interference that may occur during extreme activities. The analysis of available measurement methods, such as electrocardiography (ECG), impedance cardiography (ICG), electromyography (EMG), as well as techniques for monitoring blood pressure and blood flow in the carotid arteries, was a key step in selecting the most effective measurement techniques for the developed modules. Each of the considered methods has unique advantages and limitations that determine its usability in mobile applications, particularly in extreme environmental conditions such as those encountered in aviation [15].

Electrocardiography (ECG), as one of the most commonly used methods for monitoring heart function, offers the ability to continuously record the electrical activity of the heart muscle. It is a well-established method in both clinical and field conditions; however, its effectiveness may be limited in situations where body movements and mechanical disturbances affect signal quality [2].

Impedance cardiography (ICG) is a non-invasive method that enables the assessment of stroke volume and other hemodynamic parameters. The analysis of cyclic changes in thoracic impedance allows for the determination of parameters related to volumetric blood flow during left ventricular contraction, such as stroke volume (SV) and cardiac output (CO) [3]. Due to its non-invasiveness and ability to provide continuous monitoring, impedance cardiography is used for evaluating hemodynamic parameters both at rest and during physical exertion [4].

Electromyography (EMG), primarily used for monitoring muscle activity, can be applied in the assessment of muscle fatigue and physiological load during intense physical tasks. The analysis of EMG signals enables the quantitative evaluation of changes occurring in muscles during repetitive work related activities, which is crucial for preventing and mitigating harmful factors affecting the musculoskeletal system [5]. Studies have shown that EMG is an effective tool for monitoring muscle fatigue during intense physical tasks. The analysis of EMG signal parameters, such as amplitude and frequency, enables the assessment of muscle fatigue levels and physiological load, which is essential in both occupational ergonomics and sports training [6].

In the field of blood pressure monitoring, non-invasive methods for continuous measurement were considered. One such method is the CNAP 500 monitor, which enables continuous, non-invasive blood pressure monitoring in a beat-to-beat mode, meaning for each cardiac cycle. This device allows for the monitoring of blood pressure and heart rate for up to 24 h [7]. Another analyzed approach was the use of resonant sonomanometry, which utilizes sound waves to induce resonance in the artery and then employs ultrasound imaging to measure blood pressure. This method represents an innovative approach to non-invasive, continuous arterial blood pressure monitoring [8]. Additionally, Touchwave technology offers a groundbreaking method for continuous blood pressure monitoring without the need for traditional methods such as an inflated cuff or catheter. It is a modern approach to continuous, non-invasive arterial blood pressure measurement [9].

Particular attention was given to methods for monitoring blood flow, among which the analysis showed the superiority of Doppler technology. Doppler technology is widely used in monitoring blood flow, enabling accurate assessment of the velocity and direction of flow in arteries. It is a non-invasive diagnostic method that allows for quick and safe evaluation of the condition of blood vessels and the detection of potential abnormalities, such as stenosis or blockages [10]. Modern technological solutions allow for a significant reduction in the size of devices, making them convenient for use in mobile measurement systems. Unlike other techniques, Doppler methods demonstrate high tolerance to vibrations and movement, making them an ideal solution in dynamic conditions, such as a pilot's work during flight [11, 12]. Doppler technology, with its properties and ability to adapt to extreme conditions, is currently one of the most promising solutions for mobile monitoring of physiological status in aviation and other fields that require high precision measurements in challenging work environments.

This article presents the concept of measurement modules that enable continuous monitoring of the aforementioned physiological parameters in extreme conditions. The designed devices address the existing gap in available mobile technologies, which have not been sufficiently precise or resistant to the challenging conditions experienced during flight.

The primary aim of this study is to present a set of advanced measurement modules for real-time physiological monitoring in extreme conditions, particularly in aviation. The proposed system is designed to integrate multiple sensing technologies to ensure high measurement accuracy, mobility, and resistance to external disturbances such as movement, pressure fluctuations, and vibrations. The key innovation of the proposed approach lies in its ability to combine different monitoring techniques, ensuring a comprehensive assessment of the user's physiological state in real-time. To validate the effectiveness and reliability of the developed measurement modules, a series of controlled laboratory experiments and field tests is planned to be conducted in the future. The validation process will assess the accuracy, stability, and resistance of the system to external factors that may affect signal quality. The results obtained from these tests will demonstrate the capability of the modules to perform reliably under extreme conditions, providing crucial insights into their practical application.

This article presents the concept of measurement modules that enable continuous monitoring of the aforementioned physiological parameters in extreme conditions. The

designed devices address the existing gap in available mobile technologies, which have not been sufficiently precise or resistant to the challenging conditions experienced during flight.

2 Material and Methods

The aim of the conducted work was to develop measurement modules that form an integral part of an advanced hardware-software system designed for comprehensive acquisition, analysis, presentation, and archiving of physiological parameters. The developed system serves as a measurement station that enables the study of the human body's responses under specific experimental conditions, such as ischemic hypoxia and orthostatic stress.

Ischemic hypoxia is a condition in which tissue oxygen deprivation occurs due to restricted blood flow, which can have a significant impact on the functioning of the cardiovascular and nervous systems [13, 15]. Orthostatic stress, resulting from a sudden change in body position from lying to standing, can lead to the occurrence of orthostatic hypotension, which is a sudden drop in blood pressure. This, in turn, can cause dizziness, weakness, and, in extreme cases, even loss of consciousness [13]. The use of the developed measurement modules enables precise monitoring of key physiological parameters such as blood pressure, heart rate, blood oxygen saturation (SpO_2), and nervous system activity. The system not only allows for the real time recording of these data but also their visualization and long term analysis. This is particularly important in scientific research related to the adaptive regulatory mechanisms of the body and potential health risks associated with extreme physiological conditions.

As part of the ongoing work, a concept for a portable module for measuring blood flow in the carotid artery was developed, along with other modules for measuring physiological parameters. These include a module for assessing stroke volume using impedance cardiography (ICG), as well as modules for recording electrocardiographic (ECG), electromyographic (EMG) signals, and blood pressure waveforms, using non-invasive methods. A key aspect of the developed system is the proper placement of measurement modules on the subject's body, ensuring optimal signal quality and user comfort. To enhance ergonomics and protect the devices from external factors, dedicated housing models were designed, tailored to the specifics of each sensor.

The measurement modules were developed in accordance with the requirements of the European Parliament and Council Regulation (EU) 2017/745 on medical devices (MDR – Medical Device Regulation). Due to their compliance with MDR, the developed measurement modules will undergo a conformity assessment process, which includes several stages such as risk analysis, verification of user safety, and functional and clinical testing.

The recorders were designed in such a way that the hardware module integrating the controlling microcontroller is universal, as shown in Fig. 1. Additionally, a universal multi channel A/D converter, a wired power supply module, and a wireless charging block for the built in battery were developed, along with a data acquisition block on the built in memory card and wireless data transmission.

As part of the ongoing work, the design of a blood flow signal recorder for the carotid artery and the design of modules for measuring biomedical signals such as ECG with

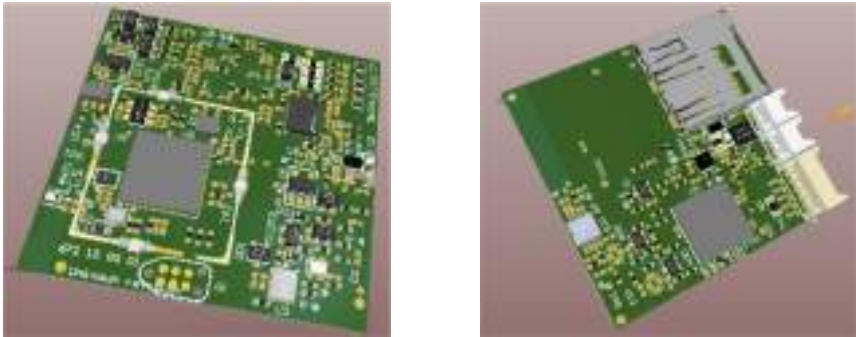


Fig. 1. View of the universal data acquisition, processing, and transmission module model, common to all measurement modules.

bioimpedance signals and EMG were developed. For all the completed models, various solutions for the construction and appearance of their enclosures, as well as different methods of attachment to the patient's body, were tested. The result of this work was the development and creation of enclosure models and the model of the system for attaching them to the patient's body, as shown in Fig. 2.

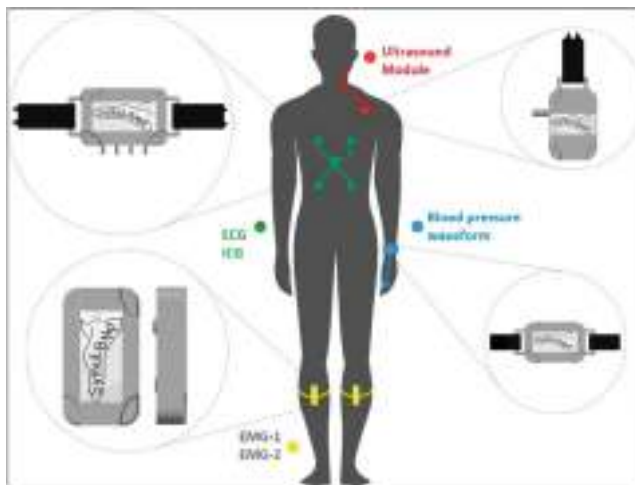


Fig. 2. Placement of measurement modules on the patient's body.

3 Results

In this section, descriptions of the developed measurement modules for the physiological parameter monitoring system will be presented. Each of these modules has been designed with the goal of precisely recording and analyzing specific biomedical signals.

Additionally, the developed data concentrator will be presented, acting as the integrator of these modules. The concentrator enables synchronization and transmission of the collected measurement data to the central system, ensuring smooth processing, analysis, and archiving.

3.1 Registration of Blood Flow in the Carotid Artery

Measurement of blood flow in the carotid artery is a key element in the physiological assessment of pilots and air force personnel, particularly in the context of exposure to extreme aviation conditions. During aviation maneuvers, such as sudden changes in g-force, significant disruptions in cerebral blood flow may occur, increasing the risk of brain hypoxia, which, in extreme cases, can lead to loss of consciousness (G-LOC, G-induced Loss of Consciousness) [13].

Monitoring of blood flow in the carotid artery can be conducted both in real flight conditions and in controlled laboratory settings, such as during tests in an overload centrifuge. Analyzing the body’s response to simulated overload conditions allows for the assessment of individual suitability for service in the air force. As part of the ongoing work, a mobile blood flow recorder model was developed, utilizing Doppler ultrasound technology. This device is innovative and original, enabling continuous, non-invasive real time monitoring of hemodynamics. By using ultrasound technology, it is possible to precisely determine key blood flow parameters such as flow velocity, cardiac output, and Doppler wave characteristics, which are important for complementing the assessment of the body’s adaptive capabilities to extreme physiological conditions. The technical specifications of the developed blood flow recorder are presented in Table 1.

Table 1. Specifications of the blood flow recorder (Ultrasound Module).

Signal	Signal connection	Parameters				Model
Carotid artery flow measured by Doppler ultrasound	- ultrasound probe attached to the patient's neck - wired connection of the probe to the signal processing module		Min	Typical	Max	
		ultrasound frequency [MHz]	2	5	8	
		ultrasonic power [MPa]			1	
		sampling rate [Hz]		100	500	
		sample size [bit]		16		

Table 1 presents the key parameters of the blood flow recorder, which uses an ultrasound probe mounted on the patient’s neck, along with an image of the assembled electronic system. The device utilizes a single measurement channel, allowing monitoring of blood flow in one vessel. The ultrasound frequency ranges from 2 to 8 MHz,

enabling adjustment of the penetration depth. The ultrasound power is up to 1 MPa, and the sampling frequency ranges from 100 to 500 Hz, allowing precise reproduction of dynamic changes in flow. The sample resolution is 16 bits, ensuring high signal quality and accurate measurements.

3.2 ECG and ICG Signal Registration

The developed measurement module is designed for simultaneous recording of electrocardiographic (ECG) and impedance cardiography (ICG) signals, enabling comprehensive cardiovascular assessment. The module employs a chest lead configuration to capture key hemodynamic parameters, including stroke volume (SV), cardiac output (CO), cardiac index (CI), left ventricular ejection time (LVET), and baseline impedance (Z0). Signal processing techniques, such as high-pass filtering to remove DC components and real time demodulation of received signals, further enhance measurement precision. The technical specifications of the ECG and ICG recorder are shown in Table 2.

Table 2. Specifications of the ECG and ICG recorder.

Signal	Signal connection	Parameters				Model
ECG and ICG signal with defibrillation protection	- standard patient cables, separate for each electrode - ADEFI patient cable socket (developed by Łukasiewicz - KIT)	ECG	Min	Typical	Max	
		sensitivity [μV/LSB]	-	2,36	2,50	
		hardware sampling rate [Hz]	500	500		
		allowable electrode potential [mV]	±300			
		ECG track range [mV]	±320			
		input impedance [MΩ]	10			
		common mode rejection [dB]	100			
		number of converter bits	18	19	24	
		frequency response [Hz]	0,05		150	
		ICG	Min	Typical	Max	
		application current amplitude [mASK]			1	
		application current frequency [kHz]	20	40	100	
		range of constant impedance component [Ω]	5		100	
		range of variable impedance component [Ω]	0		1	
		sampling rate [Hz]	500			

Table 2 presents the technical characteristics of the ECG and ICG recorder, including key parameters of both signals and equipment features related to its precision and performance, as well as the image of the assembled electronics and the enclosure model for the measurement module. The ECG recorder features a sensitivity range of 2.36–2.50 $\mu\text{V}/\text{LSB}$ and a sampling frequency of 500 Hz, ensuring accurate reproduction of heart signals. The input impedance is 10 $\text{M}\Omega$, and the common mode rejection ratio is 100 dB, minimizing interference. The bandwidth covers the range of 0.05–150 Hz, and

the converter’s bit depth ranges from 18 to 24 bits, ensuring high resolution. The ICG recorder, on the other hand, supports a single channel with a sampling frequency of 500 Hz. The applied current has an amplitude of 1 mA and a frequency of 20–100 kHz. The impedance range spans 5–100 Ω (DC component) and 0–1 Ω (AC component). The converter’s bit depth is 24 bits, ensuring precise measurements.

3.3 EMG Signal Registration

The developed EMG module is designed to record the activity of two muscle groups separately for each leg of the examined person. It enables two channel electromyographic signal acquisition with patient feedback. The system ensures proper signal quality through features such as low-pass hardware filtering, high-pass prefiltering to maintain the diagnostic EMG bandwidth, and the detection of improper electrode contact. The technical specifications of the EMG recorder are presented in Table 3.

Table 3. Specifications of the EMG recorder.

Signal	Signal connection	Parameters				Model
An electromyograph that shows the activity of two muscle groups, separately for each leg of the examined person	standard patient cables, separate for each electrode	EMG	Min	Typical	Max	
		input impedance [$M\Omega$]	10			
		common mode rejection [dB]	100			
		conversion range [μV]	1		10000	
		frequency response [Hz]	1		250	
		sampling rate [Hz]		1000	2000	
		number of converter bits		24		

Table 3 presents the detailed technical parameters of the electromyographic (EMG) recorder designed to measure the activity of two muscle groups, one on each leg of the subject. The device features an input impedance of 10 $M\Omega$, ensuring high resistance to external interference and enabling the acquisition of high quality signals. The higher impedance allows for precise measurements with minimal signal loss, which is crucial in EMG diagnostics. The recorder is equipped with a common mode rejection system of 100 dB, effectively eliminating noise and interference from other sources, such as power supply or nearby devices. The signal conversion range is from 1 μV to 10000 μV ,

allowing for the detection of both very small and large changes in muscle activity, thus ensuring a wide dynamic measurement range. The device's frequency response spans from 1 Hz to 250 Hz, enabling the recording of low and medium frequency signals typical of muscle activity. The sampling rate, ranging from 1000 to 2000 Hz, provides sufficient resolution for accurately capturing rapid changes in muscle activity, which is essential for analyzing dynamic processes in the body. These technical parameters make the EMG recorder highly precise and reliable in monitoring the electrical activity of muscles.

3.4 Blood Pressure Waveform Registration

The OEM Module (Original Equipment Manufacturer) is used for continuous non-invasive blood pressure monitoring through peripheral pressure waveforms. It provides real time registration and presentation of pressure values along with the analysis of the pressure waveform. Key parameters measured on a beat-to-beat basis include SYS (Systolic Pressure), DIA (Diastolic Pressure), MAP (Mean Arterial Pressure), HR (Heart Rate). This system utilizes a certified OEM module to ensure accurate measurements and pressure waveform analysis, focusing on the unloading of vessel walls during measurement. The description of the OEM module is included in Table 4.

Table 4. Specifications of the OEM Module (CNAP).

Signal	Signal connection	Parameters				OEM Module [14]
Peripheral pressure signal	implemented based on the OEM module	CBP	Min	Typical	Max	
		cuff pressure range [mmHg]	0		300	
		pressure measurement error [mmHg]			±5	
		Pulse Rate Range [bpm]	30		200	

The CNAP finger sensor features a simple slip on design for easy placement on two adjacent fingers, ensuring correct positioning without additional fastening. It combines infrared sensors, electronics, and inflatable pressure chambers within soft tubes for a comfortable, secure fit. The sensor automatically deflates after measurements, making removal easy [14]. Designed for long term use, the IP-protected sensor remains stable throughout measurements, with no shifting or slipping. It allows for switching the measurement finger at set intervals or manually, enabling up to 24 h of continuous monitoring. Available in three sizes, the CNAP sensor provides a personalized fit for various users [14]. The CNAP technology is based on the Vascular Unloading Technique, which originated with the Czech physiologist Jan Peňáz in the 1970s [14]. This technique forms

the basis for detecting blood volume changes in the finger and transforming plethysmographic signals into continuous blood pressure information [14]. The combination of this innovative technology with the CNAP sensor’s ergonomic design enables reliable, real time blood pressure monitoring even in extreme conditions, making it an excellent choice for long term monitoring in dynamic and challenging environments.

3.5 Data Concentrator

The role of the data concentrator is to facilitate bidirectional communication between the IT system and patient modules. As part of the conducted work, a hardware module of the medical data concentrator was developed. On the hardware level, support for two Bluetooth 5.0 Low Energy transceiver units was implemented. The adopted distribution of biomedical signal recording modules between the two BT transmission-reception channels integrated into the data concentrator is presented in the figure below, Fig. 3.

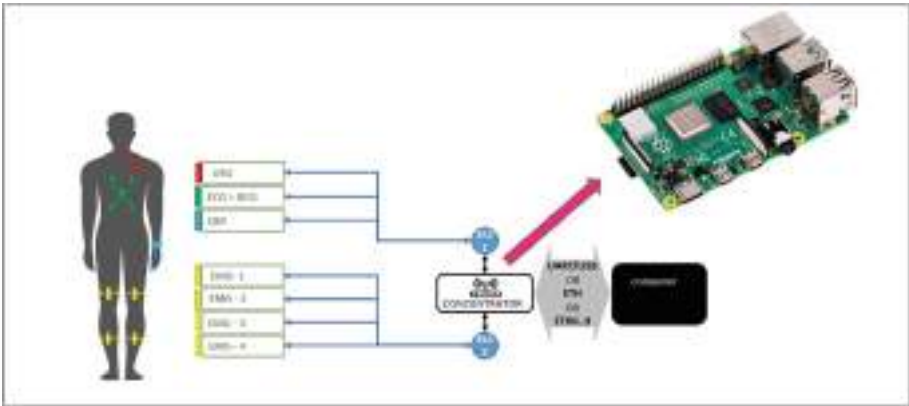


Fig. 3. Block diagram – cooperation of input modules, concentrator, and data recording computer.

The hardware module of the data terminal was developed using a Raspberry Pi microcomputer, operating under a Linux-based operating system. Dedicated data transmission algorithms were designed to facilitate communication between wireless BT links and asynchronous serial data transmission interfaces such as USB, UART, and Ethernet, allowing the terminal to connect to the data recording computer. The Raspberry Pi microcontroller module is powered either via an external power supply when using USB or UART transmission or through the Power-over-Ethernet (PoE) mechanism when using the Ethernet interface. The developed data transmission algorithms ensure bidirectional communication between patient measurement modules and the computer software. The correctness of connection establishment, data recording initiation and termination, and error signaling was verified through built in control mechanisms within the protocol, as well as dedicated test software designed to automate error detection and reporting in data transmission. Additionally, efforts were made to develop synchronization mechanisms for data transmitted from various patient modules to enable simultaneous recording and, consequently, synchronized presentation in the computer software.

4 Comparison with Existing Solutions

In this section, we compare the proposed measurement modules with existing solutions used for physiological monitoring in extreme environments, focusing on key areas such as measurement accuracy, mobile application feasibility, resistance to external interferences, and real-time data acquisition capabilities:

- **Measurement Accuracy**

Existing monitoring systems, such as electrocardiography (ECG) and impedance cardiography (ICG), are widely used for tracking cardiovascular parameters. However, conventional devices often struggle with signal interference caused by physical movement or mechanical disturbances, especially in dynamic environments like aviation. The developed modules incorporate advanced signal processing techniques to ensure higher precision even under such challenging conditions.

- **Mobility and Ergonomics**

Current systems typically lack the flexibility to operate in mobile, real-time applications without compromising performance. For instance, many ECG and ICG systems used in aviation or sports require bulky setups that limit the subject's movement. In contrast, the measurement modules developed in this study are designed to be compact, portable, and ergonomic, ensuring users can perform their tasks without restriction. This makes them particularly suitable for use in environments where continuous monitoring is critical, but mobility is equally important, such as in aviation, military, or professional sports.

- **Resistance to External Interference**

Existing non-invasive measurement technologies, such as traditional ECG and ICG, often face challenges with signal quality due to external interferences, such as vibrations or electromagnetic noise. The proposed modules are specifically engineered to minimize the impact of these disturbances. This is achieved through the integration of advanced filtering techniques and the use of highly sensitive, interference-resistant sensors, which ensures accurate data acquisition even in turbulent or high-vibration environments. This makes the proposed modules more suitable for extreme environments than many existing solutions.

- **Real-Time Data Acquisition and Integration**

While many traditional solutions focus on periodic monitoring or require post-measurement analysis, the developed system enables continuous, real-time data acquisition and integration. This is crucial in preventing potential health risks by allowing for early detection of abnormalities such as ischemic hypoxia or orthostatic stress, which may otherwise go unnoticed. The data concentrator designed for this system ensures seamless communication between the measurement modules and the central analysis system, allowing for immediate response to detected issues. This integration is a significant step forward compared to older, often isolated, monitoring systems.

• **Adaptability to Extreme Conditions**

The developed modules are tailored specifically for extreme conditions, with particular attention to adaptability in environments like high-altitude flight or intense physical activity. Unlike many existing solutions, which are primarily designed for clinical settings or stationary use, the modules can withstand the dynamic and challenging environments encountered during flight or other high-stress tasks. The integration of Doppler ultrasound, for example, enhances the ability to continuously monitor blood flow in the carotid artery during high-G maneuvers—something traditional blood flow measurement techniques cannot achieve effectively.

Table 5. Comparison of the proposed modules with existing solutions in terms of technical parameters.

Parameter	Proposed Solution	Existing Solutions
Ultrasound Module		
Ultrasound frequency [MHz]	2–8	2–5 [16]
Ultrasound power [MPa]	1	0.5–2 [17] 1 [18]
Sampling rate [Hz]	100–500	100–2000 [19]
ECG and ICG Module		
Sensitivity [$\mu V/LSB$]	2.36	2.44 [20]
Hardware sampling rate [Hz]	500	500 [21]
Input impedance [$M\Omega$]	10	> 10 [20]
Common mode rejection [dB]	100	> 90 [20]
Drive range (transferring ECG signal) [mV]	± 20	± 20 [22]
Application current amplitude ICG [mA]	1	1 [23]
Application current frequency ICG [kHz]	40	100 [23]
Sampling rate ICG [Hz]	500	500 [24]
EMG Module		
Input impedance [$M\Omega$]	10	$\geq 1 M\Omega$ [25]
Bandwidth [Hz]	250	1 Hz–10 kHz [25]

The Table 5 presents a comparison of the proposed solutions for the ultrasound (USG), electrocardiography (ECG), impedance cardiography (ICG), and electromyography (EMG) modules with existing solutions in terms of key technical parameters. The proposed ultrasound module offers a higher frequency range (2–8 MHz) and comparable ultrasound power (1 MPa) to existing solutions. However, its sampling rate (100–500 Hz) is lower than some existing solutions (up to 2000 Hz). For the ECG and ICG modules, the proposed sensitivity is slightly lower but still comparable. It has a higher input impedance (10 $M\Omega$) and better common mode rejection (100 dB) than

existing solutions. The ICG operates at a lower current frequency (40 kHz) than current solutions (100 kHz). The proposed EMG module improves input impedance (10 M Ω) over existing solutions, but its bandwidth (250 Hz) is narrower compared to existing solutions (1 Hz–10 kHz), suggesting a more specialized design.

5 Summary and Conclusions

The development of digital measurement modules for non-invasive physiological monitoring plays a crucial role in ensuring the safety and efficiency of individuals operating in extreme environments. This study presents measurement techniques, including electrocardiography (ECG), impedance cardiography (ICG), electromyography (EMG), and Doppler ultrasonography, to assess cardiovascular, respiratory, and musculoskeletal system activity. These modules are optimized for mobile applications, offering high accuracy, resistance to external interference, and real time data acquisition. Their implementation facilitates continuous health monitoring, early detection of potential hazards, and improved performance in aviation, sports, and high risk industries. The collected physiological data also contribute to research on human adaptation to extreme conditions, helping to refine safety protocols and enhance occupational and operational performance.

The presented digital measurement modules mark a significant advancement in the field of physiological monitoring under extreme conditions. Their ability to provide continuous, non-invasive, and precise data collection enhances safety and performance for individuals exposed to challenging environments. By integrating multiple sensing technologies into mobile and ergonomic designs, these modules bridge the gap between conventional clinical monitoring and real world applications. The technology complies with medical regulatory standards and undergoes rigorous validation to ensure reliability. Future developments will focus on integrating AI-driven analytics for predictive health assessments. The findings of this research pave the way for broader applications in aviation, military, sports, and other high risk professions, contributing to the overall well-being and safety of individuals working in extreme conditions.

Funding. This work was supported by the resources of the National Centre of Research and Development, Poland, within the framework of the Program for National Defense and Security, Project DOB-BIO-12-05-001-2022.

References

1. Jasiński TL (2005) Znaczenie ukierunkowanego treningu fizycznego w zwiększaniu tolerancji organizmu pilota wojskowego na przyspieszenia +Gz. AWF im. Bronisława Czecha w Krakowie, Studia i Monografie nr 32, Kraków, Warszawa
2. Augustyniak P (2001) Przetwarzanie sygnałów elektrodiagnostycznych. AGH Uczelniane Wydawnictwo Naukowe – Dydaktyczne. AGH Kraków
3. Krzesiński P, Gielerak G, Kowal J (2009) Kardiografia impedancyjna - nowoczesne narzędzie terapii monitorowanej chorób układu krążenia [Impedance cardiography - a modern tool for monitoring therapy of cardiovascular diseases]. *Kardiol Pol*

4. Krześciński, Paweł et al (2009) Impedance cardiography - A modern tool for monitoring therapy of cardiovascular diseases. *Kardiologia polska* 67:65–71
5. Procedura oceny zmęczenia mięśniowego na stanowiskach pracy powtarzalnej. https://m.ciop.pl/CIOPPortalWAR/file/91553/Procedura_oceny_2Z24.pdf?utm_source=chatgpt.com. Accessed 28 Jan 2025
6. Bartuzi P, Roman-Liu D (2007) Assessment of musculoskeletal load and fatigue with electromyography, *Bezpieczeństwo Pracy: nauka i praktyka*, vol 4. Warszawa
7. Dewhirst E, et al (2013) Accuracy of the CNAP monitor, a noninvasive continuous blood pressure device, in providing beat-to-beat blood pressure readings in the prone position. *J Clin Anesth* 25:309–313
8. Sonometria rezonansowa. <https://pulsmedycyny.pl/medycyna/kardiochirurgia/naukowcy-opracowali-nowa-metode-ciaglego-pomiaru-cisnienia-tetniczego/>. Accessed 28 Jan 2025
9. https://www.medisen sonic.com/product/touchwave?utm_source=chatgpt.com. Accessed 28 Jan 2025
10. https://bodymove.pl/usg-dopplerowskie-ocena-przeplywu-krwi/?utm_source=chatgpt.com. Accessed 28 Jan 2025
11. Badani KK, et al (2013) A pilot study of laparoscopic Doppler ultrasound probe to map arterial vascular flow within the neurovascular bundle during robot-assisted radical prostatectomy. *Prostate cancer* 2013:810715
12. Kenny JS, et al (2014) Simultaneous venous-arterial Doppler ultrasound during early fluid resuscitation to characterize a novel Doppler starling curve: a prospective observational pilot study. *J Intensive Care Med* 29:628–635
13. Sobotnicka E, et al (2018) Tests for pilots under simulated hypergravity conditions - Technological challenges and research methodology. In: 2018 25th International Conference “Mixed Design of Integrated Circuits and System” (MIXDES), Gdynia, Poland, pp 465–470. <https://doi.org/10.23919/MIXDES.2018.8436651>
14. OEM Module. <https://www.cnsystems.com/technology/cnap-blood-pressure/>. Accessed 28 Jan 2025
15. Sobotnicka E, Mocha J, Badura G, Sobotnicki A, Dziuda Ł (2025) The concept of an innovative system for monitoring physiological parameters under extreme conditions. *Electronics* 14:889
16. Boyko M et al (2018) Comparison of carotid Doppler ultrasound to other angiographic modalities in the measurement of carotid artery stenosis. *J Neuroimag* 28:683–687
17. Scuderi MM, Colletini C, Viti C, Tinti E, Marone C (2017) Evolution of shear fabric in granular fault gouge from stable sliding to stick-slip and implications for fault slip mode. *Geology* 45:731–734
18. Steinke W, Rautenberg W, Schwartz A, Hennerici M (1994) Noninvasive monitoring of internal carotid artery dissection. *Stroke* 25:998–1005. <https://doi.org/10.1161/01.str.25.5.998>
19. Wang F, et al (2021) Flexible Doppler ultrasound device for the monitoring of blood flow velocity. *Sci Adv* 7:eabi9283
20. Norav Blue ECG wireless 12-Lead resting PC-based system. <https://www.danleemmedical.com/norav-blue-ecg-wireless-12-lead-resting-pc-based-system.html>. Accessed 28 Jan 2025
21. GE Healthcare. ECG Devices. <https://www.gehealthcare.com>. Accessed 28 Jan 2025
22. Schiller: ECG200 ECG System. <https://www.schiller.ch/en>. Accessed 28 Jan 2025
23. Kassanos P (2021) Bioimpedance sensors: a tutorial. *IEEE Sens J* 21:22190–22219
24. Krishnan GH, Santhosh S, Mohandass G, Sudhakar T (2024) Non-invasive bio-impedance diagnostics: delving into signal frequency and electrode placement effects. *Biomed Pharmacol J* 17:769–778
25. McManus L, De Vito G, Lowery MM (2020) Analysis and biophysics of surface EMG for physiotherapists and kinesiologists: toward a common language with rehabilitation engineers. *Front Neurol* 11:576729



A Blood Pressure Normalization Method for Aortic Pulse Wave Velocity

Kristina Kõöts^(✉) , Kristjan Pilt , and Ivo Fridolin

Department of Health Technologies, Tallinn University of Technology, Ehitajate tee 5,
19086 Tallinn, Estonia
kristina.koots@taltech.ee

Abstract. Arterial stiffness is an independent risk factor for cardiovascular diseases (CVD). Aortic pulse wave velocity (aoPWV) is a clinically used parameter for the arterial stiffness estimation, although it is influenced by blood pressure (BP).

The aim of this study was to investigate whether and how aoPWV measurements obtained with a brachial single-cuff oscillometric device Arteriograph could be BP normalized.

An exploratory study was conducted in 26 apparently healthy subjects (25–72 years). The aoPWV measurements were carried out in the supine position and during maneuvers designed to rise arterial BP. The aoPWV values were then normalized using mean arterial blood pressure (MAP). A moderate linear relationship ($r = 0.53$, $p < 0.001$) was observed between MAP and aoPWV with individual-specific linear models describing the relationship for each subject. It was shown that the subjects with elevated arterial BP and aoPWV had a BP normalized aoPWV (aoPWV_{norm}) below 10 m/s. This indicates that the cause for an increased aoPWV (> 10 m/s) was the increased BP rather than the increased intrinsic stiffness of the arteries. Measurement of aoPWV at different BP levels is suggested to enable subject-specific BP normalization when estimating arterial stiffness.

Keywords: Arterial Stiffness · Blood Pressure · Aortic Pulse Wave Velocity · Pressure-normalized Aortic PWV

1 Introduction

Arterial stiffness is a strong predictor of cardiovascular (CV) events and mortality, even in asymptomatic individuals without a CVD diagnosis [1]. Although many techniques exist for measuring arterial stiffness, pulse wave velocity (PWV) is recognized as the gold-standard by the European Society of Hypertension [2].

PWV depends on wall thickness (h), lumen diameter (D), Young's modulus (E), and on blood density (ρ), which can be expressed using the Moens-Korteweg equation:

$$PWV = \sqrt{\frac{Eh}{D\rho}} \quad (1)$$

Although the Moens-Korteweg equation does not demonstrate a direct relationship between BP and PWV, Hughes et al. 1979 [3] have shown BP dependence on elastic modulus:

$$E(P) = E_0 \cdot e^{k \cdot P}, \quad (2)$$

where P is BP in mmHg and E_0 is Young's modulus at zero BP.

Several studies have reported the pressure dependency of PWV [4–9]. Elevated BP can seemingly increase PWV, even when the intrinsic elastic mechanical properties of the artery remain unchanged [4]. Not considering the impact of BP on PWV may result in the misinterpretation of arterial stiffness, potentially leading to inappropriate treatment decisions for the patient.

Recent studies have explored the dependency of PWV on BP and proposed methods to account for this dependency in PWV assessments. Zeiff et al. 2019 [4] studied the relationship between various single-point arterial stiffness parameters and BP on the left-hand brachial artery using ultrasound. They modulated BP by raising and lowering the arm relative to the level of the heart. Giudici et al. 2022 [5] proposed a model that determines carotid artery PWV independently of BP. Likely due to the complexity of implementing the proposed approaches in routine settings and limitations of studied cohorts, no BP normalization technique of PWV has yet been adopted into clinical practice. Although it is clinically crucial to eliminate the dependence of PWV on BP [10].

The aim of this study was to explore a possibility to normalize aPWV measurements by BP using a brachial single-cuff oscillometric device Arteriograph. Based on this hypothesis, the objectives were to confirm the dependence of aPWV, measured by the Arteriograph device, on BP, with the relationship varying individually. Secondly, create a model for normalization of aPWV by BP on healthy subjects with their mean arterial pressure (MAP). Finally, to validate the potential to normalize aPWV through PWV measurements at different BP levels using the proposed methodology.

2 Methods

The study was conducted in accordance with the Declaration of Helsinki and approved by the Tallinn Medical Research Ethics Committee of the Estonian National Institute for Health Development. The study involved 26 subjects (Male 14/Female 12) aged between 25 and 72. The mean age of the participants was 40 ± 10 years. All participants appeared healthy, none of whom had been diagnosed with hypertension, hypercholesterolemia, diabetes or any other chronic diseases. The average body mass index (BMI) of the subjects was 25 ± 4 kg/m². The subjects were asked to fast, avoid drinking coffee and smoking for 8 h before the experiment. In the beginning of the study the information about the subject's age, gender, body weight and height, and smoking status was collected and entered into the database.

In this study the aPWV was measured using the Arteriograph (TensioMed, Hungary) device, which has been validated in an earlier study [11]. The device, with a cuff placed on the right upper arm, measures firstly arterial BP using oscillometric method at the level of the arm (brachial artery). Thereafter, the cuff pressure is increased above SBP

by 35 mmHg to record pulse wave signal for 10 s. The device detects from the recorded pulse wave signal the initial pulse wave and reflected waves from the bifurcation of the arteries. However, it is assumed that the major part of the reflected wave is from the aortic bifurcation. Pulse transit time (PTT) is calculated as the return time between the initial and reflected pulse waves and the distance is determined according to the instructions of the device by measuring the distance between jugulum and pubic symphysis (JS) using tape line (see Fig. 1). It was ensured that the tape line was not following the curves of the body. The aoPWV is calculated by Arteriograph device as twice the JS distance divided by PTT.

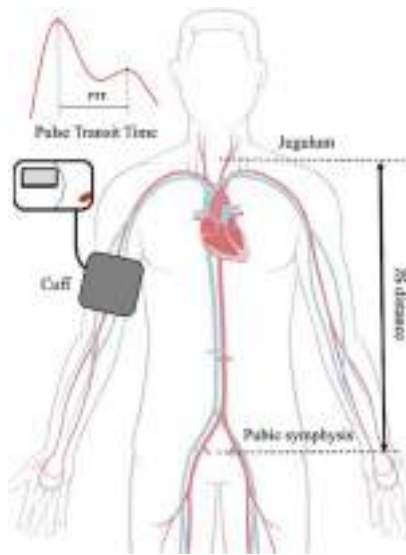


Fig. 1. Arteriograph measurement methodology

As Arteriograph device tends to overestimate BP measurements, due to its narrow cuff sizes [12], the systolic blood pressure (SBP) and diastolic blood pressure (DBP) were obtained separately using the oscillometric BP device Omron M3 – HEM – 7131 - E (Omron, Japan) at the level of the arm (brachial artery).

An overview of the study protocol is given in Fig. 2. The study measurements were performed in a quiet and temperature-controlled room (at 22 °C–23 °C). To ensure reliability of the results, the subjects were asked to rest for 10 min in the supine position before measurements. During that quiet period of time, the JS distance was measured. As follows, repeated BP measurements were obtained from the right arm using the oscillometric Omron device, as the Arteriograph measurement's methodology is designed using right arm according to the instructions. The measurement was repeated three times and the average SBP and DBP were calculated. BP measurements were performed according to the European Society of Hypertension recommendations, and the time interval between the BP measurements was not less than 1–3 min [13].

An initial aoPWV and MAP measurement was performed with the subjects lying supine, arms at the sides, and hands motionless, using the Arteriograph device. Thereafter, different physical activity maneuvers were conducted to increase BP and measure aoPWV at the different BP levels. The participant lifted one leg to a 45° angle for one minute (Fig. 2) during which aoPWV and BP were measured. After lowering the leg and resting for five minutes, aoPWV and BP were measured again.

Next, the same procedure was repeated with a 1 kg weight attached to the leg. Measurements were taken during the one-minute leg lift and again after a five-minute rest. Finally, aoPWV and BP were measured immediately after the participant performed a “leg cycling” motion in the supine position. In total, six aoPWV and BP measurements were obtained per participant.

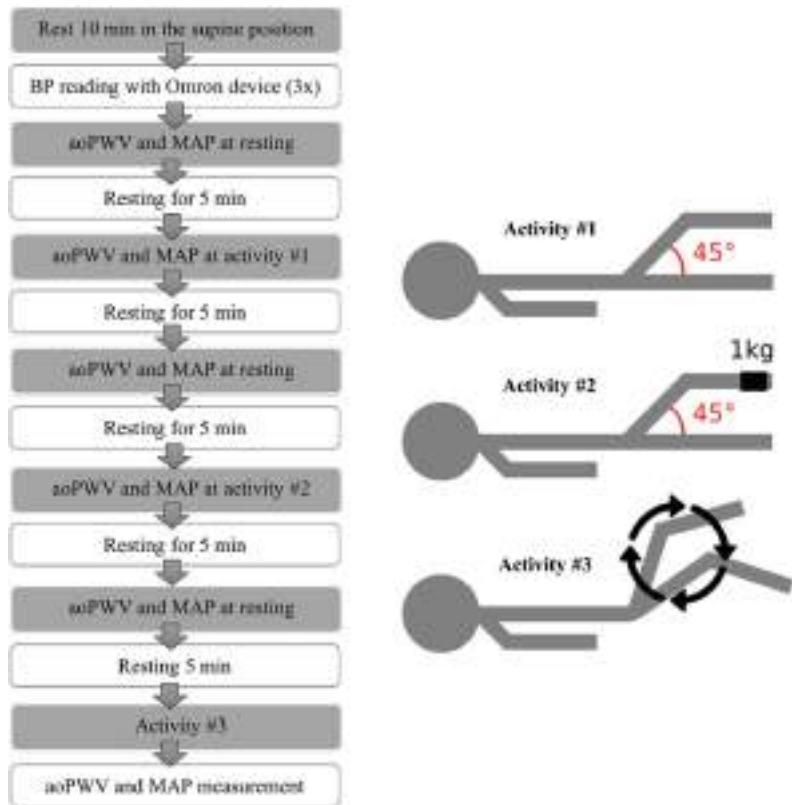


Fig. 2. Setup of the experiment and summary diagram of the methodology (Activity #1: leg up for 45°; activity #2: leg up 45° with 1 kg weight; activity #3: leg air cycling for 30 s).

3 Data Analysis and Statistical Methods

Data and statistical analysis were carried out in MATLAB (The MathWorks, USA). All results are presented as the mean $\pm$ standard deviation. The Pearson correlation coefficient was used to assess the strength and direction of the linear relationship between the variables. Linear regression models were constructed using the function `regress`. A correlation between variables with $p \leq 0.05$ was considered statistically significant.

In previous studies, PWV has been found to be related with SBP, DBP and MAP [14]. In addition, the Arteriograph utilizes the oscillometric method for BP measurement; therefore, it measures MAP while SBP and DBP are calculated using an estimation algorithm. Thus, the measured MAP provides a more accurate representation of arterial BP than SBP and DBP. In this study, aoPWV was normalized according to the MAP value.

According to the Eqs. 1 and 2, the relationship between PWV and BP is nonlinear. However, as the BP change due to the physical maneuvers is relatively small. Taylor series expansion was performed for the Moens-Korteweg equation and based on the result the linear model can be used in case of 50 mmHg MAP change. An SBP of 120 mmHg, a DBP of 80 mmHg, and consequently a MAP of 93 mmHg could be considered as normal BP. In this study, aoPWV was normalized to a MAP level of 90 mmHg.

Firstly, the linear regression model was constructed between MAP and aoPWV measured values for each subject and the model coefficients k_1 (slope) and k_2 (intercept) were calculated. Thereafter, the $aoPWV_{\text{norm}}$ value was calculated using the following equation:

$$aoPWV_{\text{norm}} = 90 \cdot k_1 + k_2, \quad (3)$$

where k_1 is the slope and k_2 is the intercept of the intrapersonal linear regression model curve, and where 90 indicates the pressure level of normalization. According to the European Society of Cardiology (ESC) guidelines the hypertension is diagnosed when SBP or DBP exceeds 140 mmHg or 90 mmHg, respectively. In addition, an aoPWV over 10 m/s is considered, according to the ESC guidelines, as a conservative estimate of significant alterations in aortic function and an increased risk of cardiovascular disease [2]. In order to show the possible clinical effect of PWV normalization the subjects were divided into four sets based on their SBP (measured with Omron device) and aoPWV values: 1) subjects with elevated SBP (≥ 140 mmHg) and aoPWV (≥ 10 m/s); 2) subjects with elevated SBP and normal aoPWV; 3) subjects with normal SBP range and elevated aoPWV; 4) subjects with normal SBP (< 140 mmHg) and normal aoPWV (< 10 m/s).

4 Results

The aoPWV measured with the Arteriograph device had a moderate linear relationship to the BP. The relationship between aoPWV and MAP measured by Arteriograph is shown in the Fig. 3a ($r = 0.52$, $p < 0.001$). The Fig. 3a includes data points of all subjects during both resting periods and physical maneuvers.

The linear regression model coefficients were calculated for each subject and the lines in Fig. 3b represent the relationship between MAP and aoPWV for each subject.

Notably, the lines exhibit different slopes and biases, confirming that the relationship between BP and aoPWV is unique for each individual.

In Table 1, the average BP, aoPWV, and aoPWV_{norm} values for each set are given. aoPWV and MAP were calculated as the average values from three-resting state aoPWV and MAP measurements obtained using the Arteriograph. SBP/DBP results are the estimated average values of three oscillometric BP measurements assessed with the Omron device. In addition, in Fig. 4 are given the average of three-resting state aoPWV and calculated aoPWV_{norm} for each subject separately.

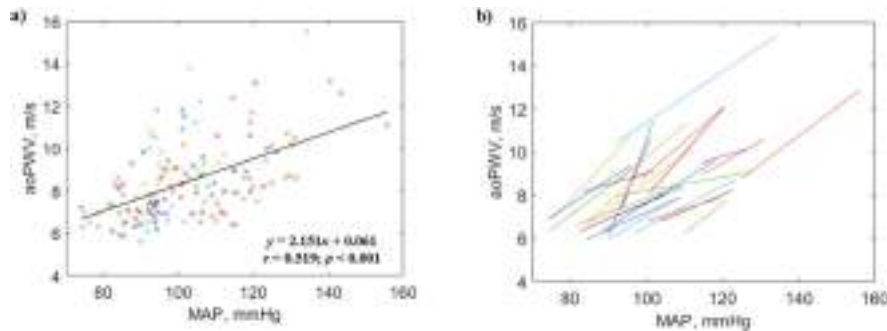


Fig. 3. The aoPWV dependence on BP. a) All subjects ($n = 26$), in a supine position, aoPWV dependence on MAP, b) All subjects' ($n = 26$) relationship between aoPWV and MAP.

Table 1. Average BP, aoPWV and aoPWV_{norm} values from three-resting states, along with standard deviations, for Sets 1–4

	Set 1	Set 2	Set 3	Set 4
	($n = 2$)	($n = 5$)	($n = 1$)	($n = 18$)
SBP/DBP	$148 \pm 3/99 \pm 1$	$147 \pm 4/88 \pm 11$	128/73	$113 \pm 28/74 \pm 7$
MAP	121 ± 4	107 ± 13	95	90 ± 8
aoPWV	10.1 ± 0.05	7.7 ± 0.70	10.9	7.0 ± 0.70
aoPWV _{norm}	7.4 ± 0.80	5.5 ± 1.70	10.3	7.3 ± 1.00
Age	45 ± 8	37 ± 4	72	39 ± 8

5 Discussion

The results from this study show that there is a relatively strong relationship between MAP and aoPWV measured by Arteriograph (Fig. 3a). Despite the scattering of results in Fig. 3a, it is attributed to individual variability in the relationship between MAP and aoPWV. Furthermore, the relationship between MAP and aoPWV is individual, and there is no universal model that can be applied to all subjects (Fig. 3b). It can be assumed that the difference between the models may be partly related to the mechanical properties of the aorta. Nevertheless, this should be confirmed in future studies.

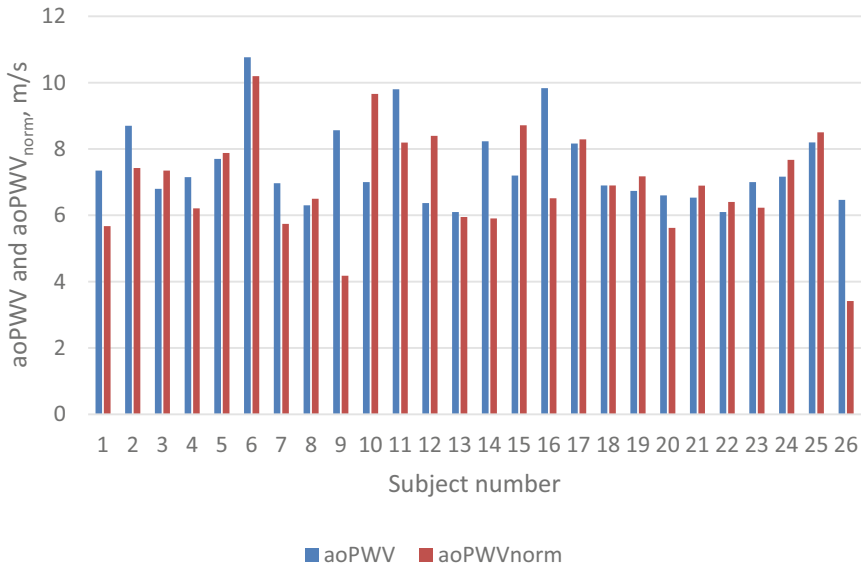


Fig. 4. The average of three resting state aoPWV and calculated aoPWV_{norm} for each subject separately.

In Fig. 4, the effect of aoPWV normalization is visible for each subject. The aoPWV_{norm} is notably decreased in subjects number 9, 16, and 26 due to high MAP values. In the case of subject number 10, the aoPWV was around 7 m/s, indicating a relatively elastic aorta. However, due to the low MAP values, it is evident that after normalization, the aoPWV_{norm} increases to nearly 10 m/s, which, according to the ESC guidelines, indicates significant alterations in aortic function.

After the implementation of MAP normalization, the aoPWV results decreased in the sets with elevated SBP. The most significant difference between aoPWV and aoPWV_{norm} was observed in the sets with elevated SBP, independent of the aoPWV values (see Table 1). After normalization in Set 1 and Set 2, aoPWV decreased by 2.7 m/s and 2.2 m/s, respectively. For the individual in the Set 3, it can be inferred that her arterial stiffness is likely elevated. This inference is supported by the fact that her PWV remains above the normal range (>10 m/s) even after BP-normalization, which is further supported by her age (see Table 1, Set 3).

Among the set of normotensive subjects with low aoPWV results, a modest increase in aoPWV was observed. This occurred because the MAP levels of certain individuals were below the threshold of 90 mmHg adopted in the applied model (see Table 1).

The results show that normalizing the aoPWV of healthy individuals based on their MAP brings the PWV down to the normal range. This approach helps determine whether the mechanical properties of the artery have actually changed. The study confirms that measuring PWV at different BP levels using the proposed methodology enables the normalization of aoPWV.

The limitation of this study is the small number of the studied subjects. Further studies with larger cohort of normotensive subjects and patients with chronic hypertension are

needed as chronic changes in BP are likely to induce changes in arterial wall mechanical properties [7, 15].

Moreover, there are studies which are critical of the brachial single-cuff methodology of PWV assessment questioning the validity of the assumption underlying the working principle of Arteriograph device, which provides only an estimation of aoPWV [16]. Nevertheless, Arteriograph has several positive characteristics and should be regarded as a viable option for noninvasive CVD risk assessment.

6 Conclusion

This study demonstrates a subject-specific dependency of aoPWV measured by the Arteriograph on BP. This provides valuable information for assessing the impact of clinical treatments or evaluating CVD risk. The effect of BP should be accounted for, leading to more accurate determination of the change in arterial wall mechanical parameters. This study proposes a simple method to eliminate the influence of BP on aoPWV results, which could potentially be applied in clinical practice for arterial stiffness evaluation using the Arteriograph device, provided it is confirmed by a further study on a larger cohort of healthy individuals and hypertensive patients.

Acknowledgements. This work was supported by the Estonian Research Council grant (PSG819).

References

1. Bonarjee VVS (2018) Arterial stiffness: a prognostic marker in coronary heart disease: available methods and clinical application. *Front Cardiovasc Med* 5:1–6. <https://doi.org/10.3389/fcvm.2018.00064>
2. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Böhm M et al (2013) 2013 ESH/ESC guidelines for the management of arterial hypertension. *Eur Heart J* 34:2159–2219. <https://doi.org/10.1093/eurheartj/ehi151>
3. Hughes DJ, Babbs CF, Geddes LA, Bourland JD (1979) Measurements of young's modulus of elasticity of the canine aorta with ultrasound. *Ultrason Imag* 1(4):356–367. <https://doi.org/10.1177/016173467900100406>
4. Zieff G et al (2019) The pressure-dependency of local measures of arterial stiffness. *J Hypertens* 37:956–963. <https://doi.org/10.1097/HJH.0000000000001998>
5. Giudici A, Palombo C, Kozakova M, Morizzo C, Cruickshank JK, Khir AW (2022) Subject-specific pressure normalization of local pulse wave velocity: separating intrinsic from acute load dependent stiffening in hypertensive patients. *Front Physiol* 12:1. <https://doi.org/10.3389/fphys.2021.783457>
6. Spronk, B., et al.: Pressure-dependence of arterial stiffness: potential clinical implications. *J Hypertens* 33:330–338 (2015). <https://doi.org/10.1097/HJH.0000000000000407>
7. Spronk B (2021) Disentangling arterial stiffness and blood pressure. *Heart Lung Circ* 30:1599–1601. <https://doi.org/10.1016/j.hlc.2021.05.086>
8. Čypienė A et al (2010) The influence of mean blood pressure on arterial stiffening and endothelial dysfunction in women with rheumatoid arthritis and systemic lupus erythematosus. *Medicina* 46(8):522–530. <https://doi.org/10.3390/medicina46080075>

9. Sezis Demirci M et al (2011) Impact of mean arterial pressure on progression of arterial stiffness in peritoneal dialysis patients under strict volume control strategy. *Clin Nephrol.* <https://doi.org/10.5414/CN107223>
10. Spronk, B., et al.: A constitutive modeling interpretation of the relationship among carotid artery stiffness, blood pressure, and age in hypertensive subjects. *Am J Physiol Heart Circ Physiol* 308:H568–H582 (2015). <https://doi.org/10.1152/ajpheart.00290.2014>
11. Horváth IG et al (2010) Invasive validation of a new oscillometric device (Arteriograph) for measuring augmentation index, central blood pressure and aortic pulse wave velocity. *J Hypertens* 28:2068–2075. <https://doi.org/10.1097/HJH.0b013e32833c8a1a>
12. Sprafka, J. M., Strickland, D., Gomez-Marin, O., Prineas, R.J.: The effect of cuff size on blood pressure measurement in adults. *Epidemiology*, 214–217 (1991). <https://doi.org/10.1097/00001648-199105000-00010>
13. 2024 European Society of Hypertension clinical practice guidelines for the management of arterial hypertension. *Eur J Internal Med* 126:1–15 (2024). <https://doi.org/10.1016/j.ejim.2024.05.033>
14. Kim EJ, Park CG, Park JS, Suh SY, Choi CU, Kim JW et al (2007) Relationship between blood pressure parameters and pulse wave velocity in normotensive and hypertensive subjects: invasive study. *J Hum Hypertens* 21:141–148. <https://doi.org/10.1038/sj.jhh.1002120>
15. Boutouyrie, P., et al.: Pharmacological modulation of arterial stiffness. *Drugs* **71**:1689–1701 (2011). <https://doi.org/10.2165/11593790-000000000-00000>
16. Segers P, Rietzschel ER, Chirinos JA (2020) How to measure arterial stiffness in humans. *Arterioscler Thromb Vasc Biol* 40:1034–1043. <https://doi.org/10.1161/ATVBAHA.119.313132>



Intelligent System for Monitoring Sedentary Behaviour to Support Elderly Care

Santiago Lynett¹, Santiago Mesa¹, Wilder Castellanos^{1(✉)}, Martha Zequera¹,
and Pilar Castellanos²

¹ Department of Electronic Engineering, Pontificia Universidad Javeriana,
Bogota, Colombia

{santiagolynett,santiagoa.mesan,wecastellanos,mzequera}@javeriana.edu.co

² School of Medicine, Universidad Pedagógica y Tecnológica de Colombia,
Tunja, Colombia

jessica.castellanos@uptc.edu.co

Abstract. The ageing population poses significant challenges to health-care systems, particularly regarding the monitoring and prevention of sedentary behaviour among the elderly. Prolonged inactivity is associated with increased risks of chronic diseases and reduced quality of life. This study aims to evaluate an innovative IoT-based intelligent monitoring system designed to track sedentary behaviour and automate attendance registration in elderly care centers. The system integrates a miniaturized and ergonomically designed wearable device with a fixed station to monitor physical activity levels and attendance in real-time. Data is transmitted to an IoT platform, providing caregivers with actionable insights and timely alerts regarding prolonged sedentary periods. A pilot study involving three elderly volunteers was conducted primarily to assess technological feasibility and the accuracy of the system in distinguishing between sedentary and non-sedentary behaviour using the Metabolic Equivalent of Task (MET) as a reference. Results indicate that the system accurately differentiates activity states; however, further studies with larger sample sizes are recommended to validate these findings comprehensively and achieve statistically significant outcomes.

Keywords: IoT · elderly care · sedentary behaviour monitoring · wearable devices · healthcare technology · MET index · real-time monitoring · RFID attendance tracking

1 Introduction

The global population is ageing at a significant rate, with projections indicating that by 2050, individuals aged 65 and older will constitute approximately 34% of the world population [1]. In Latin America and the Caribbean, the proportion of elderly is expected to increase from 9% in 2020 to 19% by 2050 [1]. Specifically,

in Colombia, the elderly population currently represents 13.9% of the total population, a figure that continues to rise consistently [2]. This demographic shift poses significant challenges for healthcare systems and elderly care institutions, as they must adapt to the growing demand for specialized care and support services.

One of the primary concerns related to aging is sedentary behaviour, which is strongly linked to various chronic health conditions, including cardiovascular diseases, diabetes, musculoskeletal disorders, and cognitive decline [3]. Studies indicate that prolonged sedentary behaviour is responsible for nearly 5 million deaths annually worldwide and contributes to an estimated \$54 billion in direct healthcare costs [4]. Older adults, particularly those in institutional care settings, are at a higher risk of prolonged inactivity due to mobility limitations, underlying health conditions, and a lack of structured physical activity programs. The health risks associated with a sedentary lifestyle underscore the need for continuous monitoring solutions that can provide real-time insights into elderly activity levels and facilitate timely interventions to promote physical movement.

In Colombia, *Centros Vida*, which are government-supported elderly care centers, provide essential services such as healthcare, nutrition, and social engagement to older adults [5]. These institutions play a crucial role in improving the quality of life for elderly individuals; however, they also face operational challenges that limit their ability to provide personalized care. A key issue is the limited availability of staff to monitor the health and activity levels of elderly residents. Caregivers often rely on manual attendance tracking and periodic health check-ups, which may lead to gaps in monitoring and an inability to detect prolonged sedentary periods in real time [6]. As a result, both elderly and caregivers experience challenges, the elderly remain at risk of health complications due to inactivity, while caregivers struggle to manage large groups with limited resources. To address these challenges, the integration of Internet of Things (IoT) technology in elderly care has gained significant attention. IoT-based monitoring solutions offer a promising approach to tracking sedentary behaviour and automating attendance registration, reducing the burden on caregivers and improving the overall quality of care.

The primary objective of this Preliminary research is to evaluate whether an IoT-based intelligent monitoring system may effectively distinguish between sedentary and non-sedentary behaviour, among elderly participants and automate attendance registration processes, thus improving care quality and operational efficiency in elderly care settings.

The system leverages wearable device and sensor-based technologies to continuously collect movement data, allowing real-time classification of activity levels and automated detection of sedentary behaviour. Additionally, cloud-based platforms enable caregivers and administrators to access activity reports, generate alerts for prolonged inactivity, and make data-driven decisions to improve elderly well-being. By implementing this technology, elderly may benefit from continuous monitoring that encourages physical activity, while caregivers gain access to automated tools that optimize their workflow and enhance care deliv-

ery. This study explores the integration and potential broader adoption of IoT solutions in elderly care settings.

This paper is structured as follows. Section 2 reviews related work on IoT systems for elderly care and monitoring of sedentary behavior. Section 3 describes the system architecture, including hardware and software components. Section 4 presents the evaluation methodology and the experimental results. Finally, Sect. 5 discusses conclusions and future research directions.

2 Related Work

IoT has revolutionized numerous industries, with healthcare being one of the most affected areas. This technology has enabled the development of innovative solutions that greatly improve the quality of life of the elderly, a rapidly growing population worldwide. Elderly care management systems that utilize IoT and computer vision techniques, as described in reference [7], provide advanced integration that enables continuous and precise elderly monitoring, thus improving the quality of care. Studies have demonstrated that the use of IoT in healthcare may reduce the burden on health systems, lower operational costs, and improve care efficiency. As highlighted in the study presented in [8], the implementation of IoT technologies in elderly care can help mitigate the impact of population aging by providing tools that allow caregivers and healthcare professionals to provide more effective and less intrusive care. These systems not only improve the quality of life for the elderly, but also enable them to maintain their independence for a longer period.

IoT systems applied to elderly care include a wide range of devices that enable the monitoring of various health conditions and behaviors. These systems may range from home-based sensors that detect environmental changes to wearable devices that track vital signs and physical activity. The study described in [9] details a fall detection system that uses IoT sensors and mobile applications to monitor the elderly in indoor environments. This system may send real-time alerts to caregivers, enabling rapid and effective intervention, which is crucial for preventing severe complications resulting from falls. Another example of IoT application in elderly care is remote health monitoring, explored in reference [10]. This research addresses the use of the Internet of Wearable Things (IoWT) for continuous monitoring of elderly. Connected wearable devices may provide real-time data on critical health parameters, such as heart rate and activity levels, facilitating continuous tracking and allowing care adjustments based on each elderly needs. This type of remote monitoring is particularly beneficial in contexts where access to healthcare services may be limited, providing an additional layer of safety and care. Two other examples of portable devices are presented in references [11, 12]. While the portable ECG device described in reference [11] offers continuous cardiac monitoring and the 3D-printed smart pill dispenser in reference [12] ensures medication adherence through IoT, both solutions have notable limitations. The portable ECG device, although accessible and efficient, often suffers from limited data accuracy due to its reliance on fewer

leads compared to hospital-grade ECG machines. Additionally, the 3D-printed smart pill dispenser, while innovative, faces challenges related to the mechanical properties and durability of the printed components. These devices may also lack the comprehensive integration and real-time alert capabilities that our system provides. Our solution not only monitors sedentary behavior and attendance but also offers a user-centered design tailored to the specific needs of caregivers and elderly users in *Centros Vida*, ensuring a more holistic and practical approach to elderly care.

In elderly care settings, IoT systems have begun to play a fundamental role in healthcare management and resident well-being. The implementation of these systems may significantly enhance the efficiency of care and the safety of residents. The elderly care management system utilizing IoT and computer vision techniques, described in reference [7], offers advanced integration with IoT and computer vision for patient monitoring. However, it has some limitations. The reliance on vision-based techniques may lead to higher implementation costs and complexity, limiting its accessibility, especially in resource-constrained environments like Latin America. Additionally, vision-based systems can be prone to errors due to lighting conditions or occlusions, impacting the reliability of the data collected.

The adoption of smartwatches among elderly, as highlighted by the authors of the study presented in [13], may enhance healthcare efficiency by providing accurate and timely health data. These devices not only assist in health monitoring but also promote autonomy and independence, which may increase overall satisfaction and well-being. Likewise, the MediFi system, presented in [14], is notable for its ability to provide portable and wireless monitoring, improving elderly health management. Wearable devices, such as smartwatches, are a crucial component of IoT applied to elderly healthcare. These devices enable continuous monitoring of vital parameters and offer additional functionalities that enhance the quality of life of the elderly. Given these advantages of wearable devices, it was adopted a wearable device-based system in our solution to ensure continuous and reliable monitoring of sedentary behaviour and attendance in *Centros Vida*. Reference [15] discusses the various ways in which wearable devices and IoT may enable assistive technologies to provide personalized support and improving autonomy of elderly. Smartwatches, for instance, may monitor heart rate, sleep patterns, and physical activity, offering data that is essential for health management. Beyond basic monitoring, wearable devices may function as personal trainers, helping elderly remain active and healthy. Reference [16] explores a low-cost indoor health monitoring system that uses wearable devices to collect real-time health data. This system offers a scalable and potentially ubiquitous solution for indoor remote health monitoring using low-energy Bluetooth beacons and a smartwatch. However, it does not emphasize real-time alerts for prolonged sedentary periods, which is a critical feature in our solution to enable timely interventions by caregivers.

The elderly interaction is fundamental to ensuring the effective adoption of these technologies, as highlighted in reference [13]. Studies on smartwatch

adoption among elderly users emphasize that usability, perceived utility, and technological familiarity are key determinants of acceptance [13]. Many older adults face cognitive and motor limitations that make traditional interfaces difficult to use, leading to low engagement with digital health solutions. To address these challenges, research suggests that interfaces should incorporate large, high-contrast text, simple navigation schemes, and haptic feedback mechanisms [15]. Additionally, ensuring minimal learning curves through adaptive tutorials or caregiver-assisted onboarding processes can significantly improve adoption rates. Our solution integrates these principles by developing an intuitive and ergonomic system tailored to the specific needs of both caregivers and elderly users, promoting a seamless and user-friendly experience. Unlike previous studies that focus solely on technical feasibility, our approach considers human factors engineering, ensuring that usability and accessibility are central to the design process.

Lastly, the psychological and social impact of wearable devices and IoT systems in elderly care should not be overlooked. The study in reference [16] examines how these technologies influence self-esteem, independence, and social interaction among older adults. Research indicates that when elderly users actively engage with these systems, they experience a greater sense of autonomy and reduced anxiety about their health status [10]. Additionally, real-time health monitoring fosters a stronger connection between patients and caregivers, reinforcing a continuous support network that enhances emotional well-being. However, poorly designed systems can have negative effects, such as increased frustration, feelings of surveillance, or resistance to adoption [9]. To mitigate these risks, studies recommend incorporating customizable alert settings, social features for peer interaction, and unobtrusive monitoring techniques [12]. Our approach aligns with these findings by ensuring that the system not only improves physical health outcomes but also enhances psychological comfort and social well-being, providing elderly users with a sense of security, empowerment, and continuous companionship.

3 Materials and Methods

3.1 MET Computation and Its Role in Sedentary Behaviour Monitoring

An important element in the developed system is the calculation of the MET index, as this parameter was used to identify if an individual is exhibiting sedentary behaviour. The Metabolic Equivalent of Task (MET) is a widely recognized unit used to quantify physical activity intensity. This metric represents energy expenditure relative to resting conditions, providing a standardized approach to classifying physical activity levels. As shown in Table 1, each physical activity for older adults is associated with an energy expenditure measured in METs [17]. Sedentary behaviour is formally defined according to [18] "*Sedentarism, sedentary behaviour, or sedentary conduct is defined as any waking behaviour*

characterized by an energy expenditure of ≤ 1.5 METs while in a sitting, reclining, or lying posture. Consequently, it must fulfill three criteria: a) it occurs during waking hours; b) it involves low energy expenditure (≤ 1.5 METs); and c) it is performed in a sitting, reclining, or lying position.” One of the methods for measuring METs is through the use of a triaxial accelerometer. These devices measure accelerations in units of meters per second squared (m/s^2) or in terms of gravitational force (g) along different axes, commonly identified as a_x , a_y , and a_z , capturing a wide range of body movements in various directions and dimensions. This facilitates a detailed and comprehensive assessment of physical activity.

Table 1. Metabolic Activity Index of Different Physical and Sports Activities in Individuals Over 64 Years Olds [17].

Intensity of Activities (for adults aged ≥ 65 years)	METs	Types of Physical Activities	Mets $\cdot min^{-1}$ (in 10 min)
Light (1.6 to 3.1 METs)	2.0	Walking around the house	20
	2.5	Walking slowly at 3.2 km/h (flat). Yoga. Stretching	25
	2.8	Walking at 4 km/h (flat)	28
	3.0	Slow ballroom dancing; foxtrot, tango, and mambo. Stationary bicycle with 50-watt resistance. Walking the dog	30
Moderate (3.2 to 4.7 METs)	3.3	Walking briskly at 4.8 km/h (flat)	33
	3.8	Walking for exercise at 5.6 km/h (flat)	38
	4.0	Tai Chi. Lelderlysure cycling (<16 km/h) (flat)	40
	4.5	Modern dance, disco, fast ballroom dancing. Golf walking and carrying clubs. Badminton singles and doubles (non-competitive)	45
Vigorous (4.8–6.7 METs)	5.0	Walking very briskly at 6.4 km/h (flat). Tennis doubles (non-competitive)	50
	5.5	Stationary bicycle (100-watt resistance)	55
	6.0	Swimming (for lelderlysure) in general. Cycling at a light pace (16–19 km/h). Dancing (non-competitive)	60
	6.3	Walking at 7.2 km/h (flat)	63

Note: 1 MET = 1 basal metabolic unit = $3.5ml \cdot kg^{-1} \cdot min^{-1}$ or $1kcal \cdot kg^{-1} \cdot h^{-1}$

The study presented in [19] defines Eq. 1 as the variable VM (Vector Magnitude), calculated as the square root of the sum of the squares of the acceleration values on the x, y, and z axes. In this study the VM parameter is related to the MET, such that if the VM value is less than 1, it corresponds to a MET of 1.5, a value that is classified as sedentary behaviour.

$$VM = \sqrt{(a_x)^2 + (a_y)^2 + (a_z)^2} \tag{1}$$

3.2 System Architecture

The architecture of the implemented system is shown in Fig. 1. The architecture is composed of four main subsystems. Firstly, the Data Acquisition Subsystem

is composed of two primary devices: a Wearable Device and a Fixed Station. The wearable device features an ESP-32 microcontroller, an OLED display, a BMA400 triaxial accelerometer for detecting accelerations in the X, Y, and Z axes, a real-time clock (RTC), and a Wi-Fi module for connectivity. The device, powered by a battery, continuously monitors the physical activity levels of the elderly. Meanwhile, the fixed station, also powered by an ESP-32 microcontroller, includes RFID sensors for user identification and three LED indicators for status display. The fixed station is used to detect the presence of the user through RFID technology, ensuring the automated registration of the elderly attendance at the elderly center. Secondly, the Connectivity Subsystem ensures the secure and reliable transmission of data acquired by the wearable device and the fixed station over a Wi-Fi network using TLS/MQTT protocols. The data is sent to cloud services, where it is stored and processed for further analysis. Next, the Data Analytics and Visualization Subsystem utilizes advanced data analytics in the cloud-based system to process the incoming data. The processed data is then visualized on a web platform accessible by healthcare staff and authorized family members. The web platform displays critical information such as activity levels, attendance records at the elderly center, and health-related notes for each elderly. Additionally, it may generate alerts for sedentary behaviour and panic button activations, ensuring timely interventions. Finally, the User Interface Subsystem provides multiple user interfaces to facilitate interaction and data access. These include the web platform for comprehensive data visualization and analysis, and Telegram integration for real-time notifications and alerts to healthcare staff and family members. The Application and Impact of the IoT system, which is implemented in Centro Vida, continuously monitors the physical activity and attendance of elderly residents. The system ability to track and analyse sedentary behaviour in real time allows caregivers to promptly address inactivity issues,

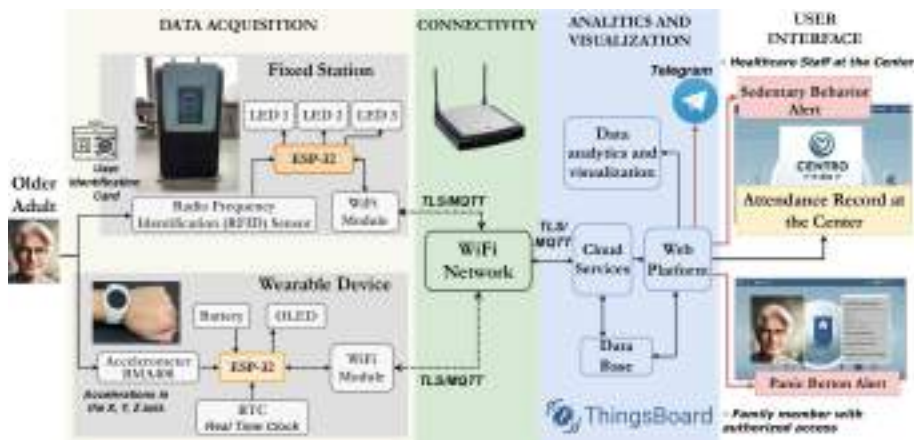


Fig. 1. Architecture of the IoT System.

thereby improving the overall quality of care and reducing the risks associated with prolonged sedentary periods.

3.3 Wearable Device

The wearable device is a lightweight, wrist-worn unit designed for continuous movement monitoring and attendance tracking. Figure 2 illustrates its key components, physical assembly, and final design.

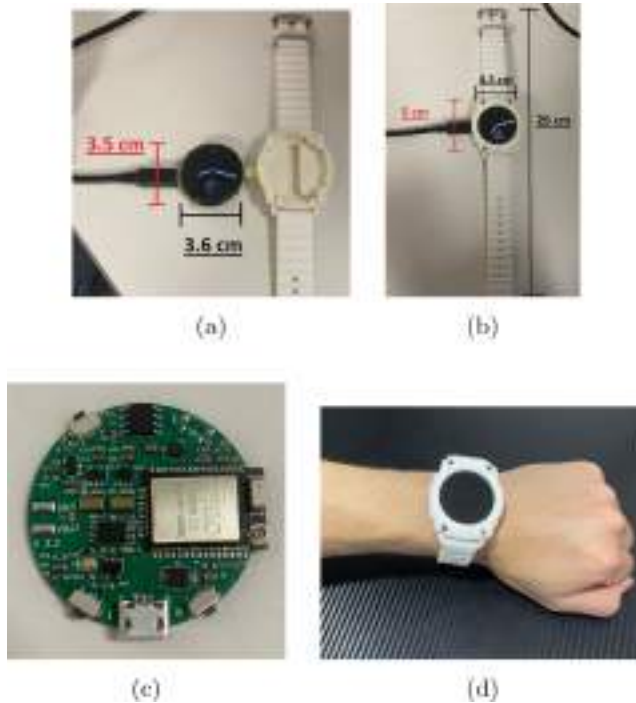


Fig. 2. Design, assembly, and final mounting of the wearable device. (a) Partially assembled wearable device, (b) Fully assembled wearable device, (c) Assembled and soldered PCB of the wearable device, (d) Wearable device on the wrist.

The wearable device is designed to perform two essential functions: Sedentary behaviour Detection and Automated Attendance Registration. It continuously calculates MET values every minute based on acceleration data obtained from the accelerometer sensor. The device is powered by an ESP32 PICO-D4 micro-controller, which processes sensor data and facilitates Wi-Fi communication with the cloud. Additionally, a 1.3-inch LCD screen, which displays essential information such as time, Wi-Fi connectivity status and prolonged periods of inactivity alerts. The wearable also interacts with the fixed station through an integrated

RFID tag, enabling seamless attendance tracking for elderly residents at *Centros Vida*.

3.4 Fixed Station

The fixed station ensures automated attendance monitoring. Figure 3 illustrates its design, including the 3D model, assembly process, and final implementation. The station is designed to operate autonomously and provide real-time feedback on attendance registration.

Hardware components include an RFID sensor, responsible for detecting the wearable device upon proximity, and an ESP32 DevKit microcontroller, which processes RFID data and transmits it to the cloud-based platform. Additionally, LED indicators provide real-time visual feedback on attendance registration status, and a protective enclosure ensures durability.

When an elderly approaches the fixed station within a 10 cm range, similar to how passengers tap their transit cards on a reader to validate their trip. The RFID module detects the wearable device's unique identifier, logs the event, and transmits attendance data to the cloud. This system eliminates the need for manual attendance tracking, enhancing operational efficiency.

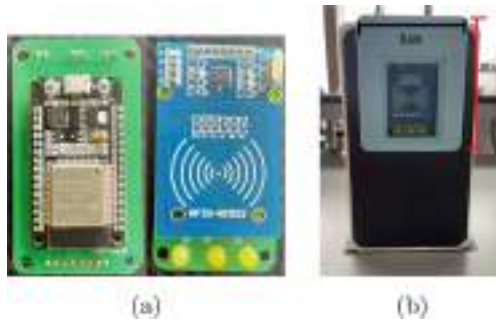


Fig. 3. Design, assembly, and mounting of the fixed station. (a) Assembled and soldered PCB circuit, (b) Final assembly of the fixed station.

3.5 Cloud Infrastructure and Data Processing

The wearable device and fixed station communicate with the ThingsBoard IoT platform via the Message Queuing Telemetry Transport (MQTT) protocol, allowing for real-time telemetry data transmission. ThingsBoard functions as the system data reception, processing, and visualization of devices-generated telemetry.

ThingsBoard utilizes message queue and rule chains to process incoming data streams. These mechanisms allow real-time filtering of data, classification

of activity levels, and the generation of alerts when predefined thresholds are met. Additionally, ThingsBoard integrates with external APIs such as Telegram to send emergency notifications to caregivers in response to critical events like extended sedentary periods or the activation of the emergency button on the wearable device, as it may be observed in Fig. 4(a), the visualization on a mobile device simulates what a caregiver at the center would see.

The platform supports role-based access control, allowing different user groups (caregivers, administrators, elderlies) to securely access system data. Moreover, the system is capable of displaying the collected data from alerts generated over one-month periods as it may observe in Fig. 4(b), enabling health-care professionals to study behavioural patterns over time and make data-driven decisions for elderly care management. the functionalities of the caregivers panel, which is shown in Fig. 4(b), caregivers will have access to the list of all active wearable devices in the Centro Vida network. Alongside each device, basic user information is displayed, including identification details, required medications, and real-time alerts, as well as historical data over time.

The administrators panel, shown in Fig. 4(c), grants access to unprocessed sensor data received from the wearable devices connected to the Centro Vida platform. This panel enables direct comparison of MET values transmitted by each wearable device against the reference MET table from the study [18], ensuring real-time assessment of user activity levels. Additionally, the administrators panel provides live sensor diagnostics, displaying whether the wearables are actively transmitting acceleration data from the onboard accelerometer sensor. This ensures continuous monitoring of device functionality and connectivity status. Furthermore, the system logs and visualizes all users who have entered the facility through the fixed RFID station, facilitating automated attendance tracking and improving operational efficiency.

Furthermore, the administration panel provides access to all dashboards available in the center. An example of the different user profiles may be seen in Fig. 4(d), where the profile of Elderly User 1 is displayed. This profile includes relevant information such as comorbidities, required medications, and pending medical appointments and all the alert history of prolonged sedentary periods or emergency button activations, to enhance timely interventions.

3.6 Controlled Laboratory Setting

The evaluation of the proposed system was conducted in a controlled laboratory setting at the research laboratory “FootLab” of Pontificia Universidad Javeriana. This environment was carefully arranged to ensure consistent testing conditions, eliminating external factors and distractions that could affect participant activity levels (See Fig. 5). The study involved a thorough examination of the system functionality, integration, and interoperability of its components. Additionally, the accuracy of the MET index calculation using the VM (vector magnitude of acceleration) was assessed. The study recruited three elderly participants, all of whom provided informed consent before participating. The inclusion crite-

ria required that participants with independent mobility and free from severe medical conditions that could influence their movement behaviour.

3.7 Experimental Protocol

The experiment followed a structured protocol designed to assess the accuracy and efficiency of the wearable device in detecting and classifying sedentary and non-sedentary behaviours. Each participant performed two primary activities, under supervised conditions:

- Seated Drawing Activity (30 min): Participants remained seated at a desk while engaging in a simple drawing task. This phase was designed to establish a baseline for sedentary behaviour, allowing the wearable device to capture low-movement periods and generate corresponding MET values.

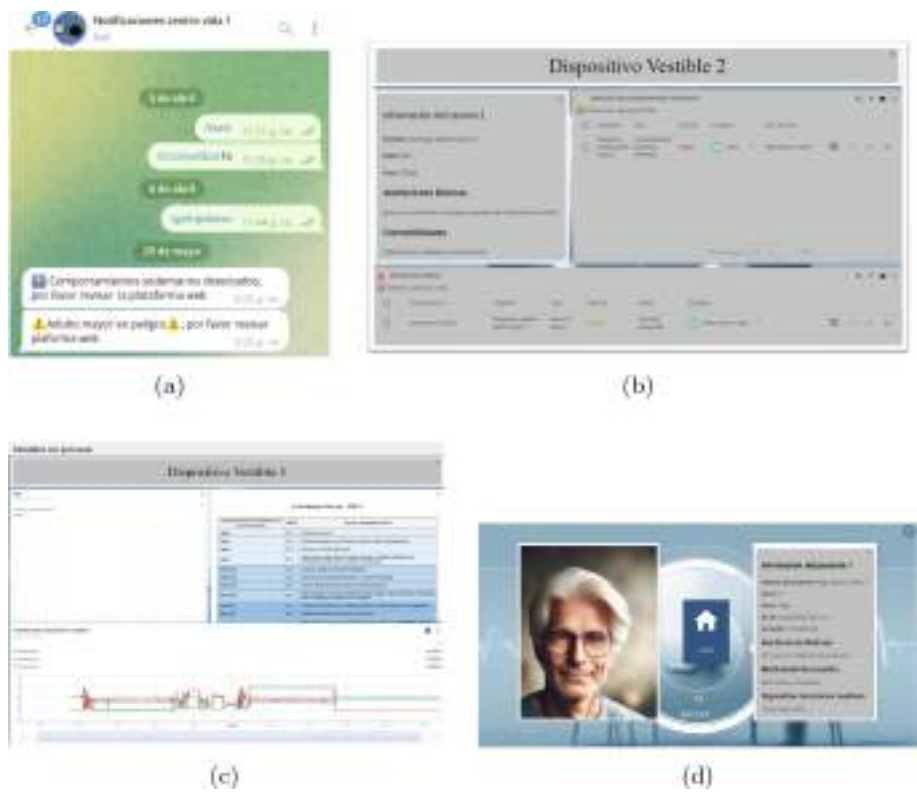


Fig. 4. Centro Vida system dashboard views. (a) Emergency Telegram notifications, (b) Alerts generated in caregivers panel, (c) Administrator panel displaying unprocessed telemetry data and MET table values, (d) Elderly User 1 panel displaying basic information and historical data.

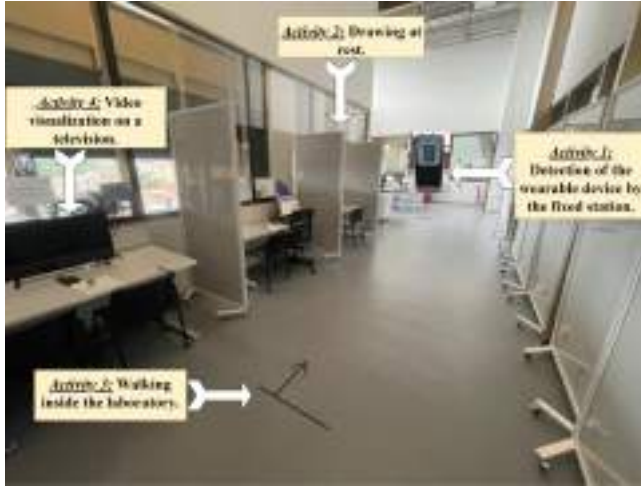


Fig. 5. Controlled environment for the tests conducted with the test subjects.

- Walking Task (10 min): Participants walked at a steady pace within the designated testing area. This activity evaluated the system ability to differentiate between sedentary and active behaviour, ensuring accurate movement classification.

Each participant underwent a 10-minute stabilization period before each activity to ensure consistency in the initial data. The wearable device continuously measured acceleration and calculated the Magnitude Vector (VM). A VM value of 1 was set as the threshold, corresponding to 1.5 MET. Values above 1 ($VM > 1$) indicate $MET > 1.5$, representing non-sedentary behaviour, while values below 1 ($VM < 1$) correspond to $MET < 1.5$, indicating sedentary behaviour. The data is presented in VM, as this is the unit measured directly by the device.

Finally, data from the wearable device were transmitted to a cloud-based database using the IoT platform ThingsBoard. The collected data were processed using the own watch script, which analysed movement patterns and categorized participant activity levels.

3.8 Evaluation Protocol and Safety Considerations

To accurately determine whether an individual exhibits sedentary behaviour, the evaluation protocol was structured into 30-minute blocks, each divided into three 10-minute segments. The system assesses the participant's physical activity level by monitoring the vector magnitude (VM). If, within at least one of these 10-minute segments, the participant surpasses the VM threshold of 1, corresponding to a MET value greater than 1.5, the system classifies the full 30-minute block as non-sedentary and does not trigger an alert. Conversely, if during all three

segments the VM remains below 1 ($MET < 1.5$), the system flags the block as sedentary behaviour and generates an alert on the ThingsBoard platform while also notifying the family members via Telegram. This structured approach ensures that transient movements do not lead to misclassification, while capturing prolonged periods of inactivity.

4 Evaluation and Results

4.1 Application of the Protocol in the Experimental Evaluation

During the experimental evaluation with elderly volunteers, the application of this protocol was adjusted for participant safety to minimize the risk of fatigue or injury. Specifically, for the walking test, participants performed the activity for only one 10-minute segment, rather than the full 30-minute block. This ensured that the system correctly detected a VM greater than 1 ($MET > 1.5$), preventing an unnecessary sedentary behaviour alert while reducing physical strain on the volunteers. In contrast, the sedentary behaviour test was conducted for the full 30 min (three 10-minute segments) to verify that the system correctly identified prolonged inactivity and triggered the appropriate alerts on ThingsBoard and Telegram. Additionally, the drawing task served a dual purpose: first, to confirm that the system accurately classified the activity as sedentary; and second, to evaluate whether wrist movements associated with colouring could erroneously be interpreted as physical activity. This aspect was crucial, as the system needed to ensure that minor hand movements did not cause a false classification of $MET > 1.5$, thereby maintaining the integrity of the sedentary behaviour detection algorithm.

4.2 Evaluation Results

The evaluation objectives for the system were threefold. Firstly, to determine the consistency of MET values obtained throughout the duration of each activity. Secondly, to observe the accuracy in recognizing sedentary and non-sedentary activities. Thirdly, to establish the influence of the hand (dominant or non-dominant) used to wear the wearable device on the MET measurements

As shown in Fig. 6, throughout the entire drawing activity, the VM values obtained remained below the threshold value of 1 for all test subjects. This demonstrates consistency and stability in the measurements obtained. The fact that the wearable device calculated a VM of 1 means that the MET is less than 1.5, corresponding to the identification of sedentary behaviour. This indicates that the system detected sedentary behaviour for all users, regardless of which hand the subject wore the wearable device on.

4.3 Evaluation Results

The evaluation objectives for the system were threefold. Firstly, to determine the consistency of MET values obtained throughout the duration of each activity.

Secondly, to observe the accuracy in recognizing sedentary and non-sedentary activities. Thirdly, to establish the influence of the hand (dominant or non-dominant) used to wear the wearable device on the MET measurements.

As shown in Fig. 6, throughout the entire drawing activity, the VM values obtained remained below the threshold value of 1 for all test subjects. This demonstrates consistency and stability in the measurements obtained. The fact that the wearable device calculated a VM of 1 means that the MET is less than 1.5, corresponding to the identification of sedentary behaviour. This indicates that the system detected sedentary behaviour for all users, regardless of which hand the subject wore the wearable device on.

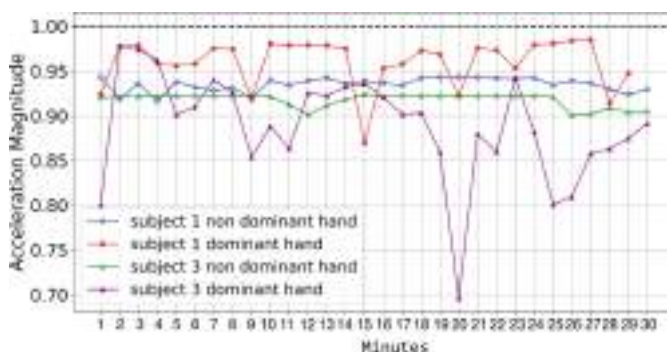


Fig. 6. Triaxial acceleration magnitude (VM) measurements during a 30-Minute sedentary activity for three test subjects using a wearable device

It should be noted that in some figures, such as Fig. 6, the data from Subject 2 is missing. This is because, on the second day of the study, Subject 2 was unable to attend the evaluation sessions due to a **family-related emergency**. As a result, their data was not recorded for that day, leading to its absence in certain graphical representations. However, the overall conclusions remain valid, as the data collected from the other participants provided **sufficient consistency** to assess the system's performance effectively.

If the VM values obtained are analyzed in more detail according to the hand each individual used to wear the wearable device, some differences may be observed (See Fig. 7). In particular, when the wearable device was used on the dominant hand, the VM values were closer to the threshold value ($VM=1$), which can be explained by the fact that this hand was used for the drawing activity. Nonetheless, the system continued to correctly recognize the activity as sedentary behaviour.

On the other hand, the results of the walking test can be observed in Figs. 8 and 9. For this activity, the VM values calculated by the wearable device identified non-sedentary behaviour in 3 out of 5 cases. For example, for subject 1, the activity was correctly identified as non-sedentary ($VM > 1$), regardless of which hand the wearable device was worn on. However, a greater dispersion of

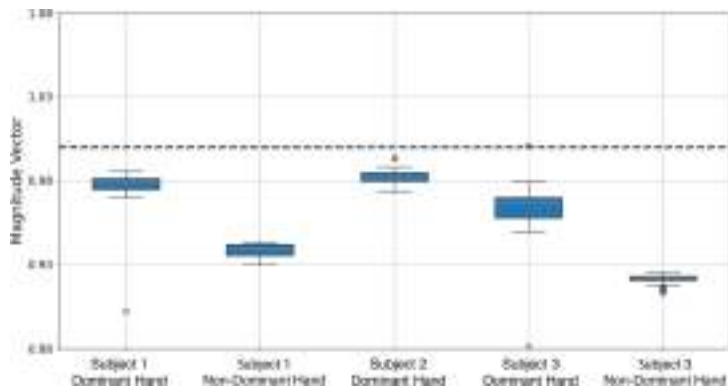


Fig. 7. Comparison of VM values obtained with the dominant and non-dominant hand during sedentary activity

data was observed compared to the other subjects. In contrast, with subject 3, less precise results were obtained; the system correctly identified non-sedentary behaviour when the wearable device was used on the non-dominant hand. But when the dominant hand was used, the system erroneously detected the activity as sedentary. The same result was observed with subject 2.

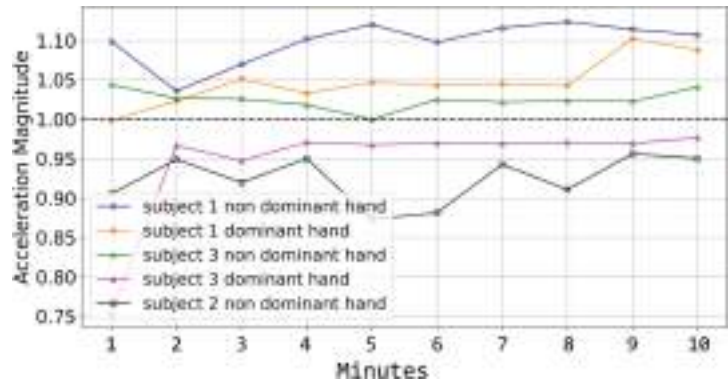


Fig. 8. Non-sedentary activity: triaxial acceleration magnitude (VM) measurements during a 30-Minute

Finally, based on the preliminary tests conducted so far, it can be concluded that more consistent results could be obtained when the non-dominant hand is used to wear the wearable device. This could improve the accuracy in identifying the type of activity. Nevertheless, more exhaustive tests must be conducted.

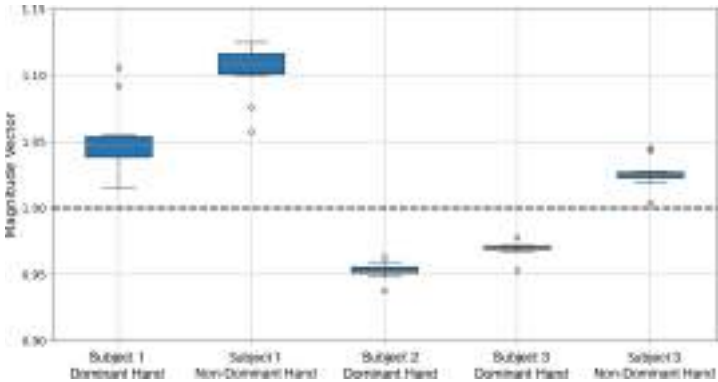


Fig. 9. Non-sedentary activity: Comparison of VM values obtained with the dominant and non-dominant hand

5 Discussion

It was reported that there are several existing solutions in the state-of-the-art that leverage accelerometers, RFID, and IoT for monitoring the activity of senior citizens. However, our study addresses a specific and underexplored context: the unique needs of Centros Vida in Colombia, day-care centers for elderly individuals in vulnerable socio-economic situations. Unlike many solutions in the state of the art that are either limited to controlled laboratory environments or designed for broader audiences, our system was developed in close collaboration with caregivers and medical staff to ensure it directly addresses the real-world challenges faced in these centers.

One of the key innovative aspects of our solution lies in the integration of the Metabolic Equivalent of Task (MET) index as a metric for monitoring sedentary behavior. While the state-of-the-art often relies on generalized methods for physical activity measurement, our approach is distinct in that it adapts and personalizes the MET calculation within the wearable device. Using a triaxial accelerometer, the device computes the Vector Magnitude (VM), enabling it to classify activity levels specific to the elderly population in Centros Vida. This customization ensures the system ability to provide accurate insights into sedentary and non-sedentary behaviors, making it uniquely suited to the needs of this context.

The tests conducted in a controlled environment confirmed the device accuracy and reliability. In terms of physical activity detection and attendance tracking via RFID technology, the device exhibited a minimal error rate. This outcome is crucial, as it ensures that the device may effectively distinguish between sedentary and non-sedentary activities, providing valuable insights into the frequency with which the elderly engage in sedentary behaviours throughout the day. Additionally, these data are essential for conducting a detailed analysis of user behaviour. The ability to store and analyse these data allows caregivers and

facility administrators to monitor trends over time, identify concerning patterns, and adapt care plans accordingly.

Additionally, our user-centered design emphasizes cost-effectiveness and practicality, addressing one of the common limitations of state-of-the-art solutions, which often result in high-cost products that are not accessible in regions like Latin America. By focusing on real-time monitoring, intuitive interfaces tailored to caregiver needs, and affordable implementation, our system strikes a balance between innovation and accessibility. This makes it a practical and impactful tool for improving elderly care.

6 Conclusions

In relation to the development of the wearable device, it is important to highlight the complexity due to the restrictions regarding energy, processing capacity and size of this type of device. In this sense, the construction of the wearable device presented significant challenges due to its compact size. Designing a device that integrates multiple components, such as an accelerometer, battery, Wi-Fi module, and microcontroller, within a limited space required careful design and high precision construction techniques. Additionally, various design and functionality aspects mentioned by older adults during the preliminary stages were also considered in the final implementation. Such as very bright lights, reduction of sound alerts, as well as avoiding the use of haptic feedback (vibration patterns).

The communication between the wearable device and the cloud platform has proven to be stable. The tests validated that the system maintained a consistent connection, ensuring the accuracy and reliability of the transmitted data at all times. This communication system is essential for continuous and real-time monitoring of the older adults. Real-time data transmission allows caregivers and healthcare personnel to access up-to-date information on the activity status of users. This continuous monitoring capability facilitates informed and timely decision-making based on accurate and current data, significantly improving the quality of care provided. Furthermore, the stability of the communication system ensures that critical data is not lost, which is fundamental for maintaining a reliable and comprehensive record of user behaviour. This not only aids in the daily management of care but also provides a robust database for future analysis and improvements in care programs.

The initial testing of the system successfully verified its functionality across all stages, ensuring correct integration and interoperability of its components. Furthermore, the tests confirmed that calculating the MET index using the VM (vector magnitude of acceleration) is a promising alternative for future studies. However, a more comprehensive evaluation, involving a larger number of participants, should be performed to further validate the effectiveness of the system in detecting sedentary behaviour. During these controlled evaluations, tests conducted with adult volunteers confirmed that the measurements obtained from the wearable device remain reliable when transmitted to the cloud platform. The characterization of the magnitude vector for sedentary and non-sedentary activities enabled an accurate classification of user movements. This classification

was based on distinct activities, such as drawing while at rest and walking at a normal pace, allowing the device to effectively differentiate between low- and high-intensity activities. Additionally, a stronger response was observed when the device was worn on the non-dominant hand, as it generated a higher magnitude vector compared to the dominant hand. This classification capability is essential for providing precise and meaningful insights into physical activity levels. The results obtained underscore the robustness of the system and its ability to operate effectively in real world settings. These findings suggest strong potential for its deployment in *Centros Vida* and similar environments, where accurate monitoring of sedentary behaviour and attendance is crucial for improving the quality of care for elderly. The tests confirmed that the system is not only a reliable solution for tracking the health and well being of this vulnerable population, but may also be seamlessly integrated into the operational workflow of *Centros Vida*, enabling more proactive and personalized care.

References

1. United nations department of economic and population division social affairs (2022). World Population Prospects 2022 Summary of Results. https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf (visited on 08/10/2023)
2. National Administrative Department of Statistics - DANE - Colombia (2019). Personas mayores en Colombia, hacia la inclusión y la participación XV Congreso Internacional de Envejecimiento y Vejez: Década del Envejecimiento Saludable 2020-2030. <https://www.dane.gov.co/files/investigaciones/notas-estadisticas/nov-2021-nota-estadistica-personas-mayores-en-colombia-presentacion.pdf> (visited on 08/10/2023)
3. World Health Organization WHO (2020): Every move counts towards better health. [www.who.int. https://www.who.int/es/news/item/25-11-2020-every-move-counts-towards-better-health-says-who](https://www.who.int/es/news/item/25-11-2020-every-move-counts-towards-better-health-says-who)
4. World Health Organization WHO (2020): Decade of healthy ageing: plan of action 2020-2030. [www.who.int. https://www.who.int/es/publications/m/item/decade-of-healthy-ageing-plan-of-action?sfrsn=b4b75ebc.25](https://www.who.int/es/publications/m/item/decade-of-healthy-ageing-plan-of-action?sfrsn=b4b75ebc.25)
5. Government of Colombia (2009): Law 1276 of 2009 new criteria for comprehensive care for the elderly in Centros Vida-Colombia. Gestor normativo. <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=34495> (visited on 08/01/2023)
6. Tremblay MS et al (2017) Sedentary Behavior Research Network (SBRN) – Terminology Consensus Project process and outcome. *Int J Behav Nut Phys Activity* 14. <https://doi.org/10.1186/s12966-017-0525-8>, <https://link.springer.com/article/10.1186/s12966-017-0525-8> (visited on 11/14/2023)
7. Paulauskaite-Taraseviciene, A, et al (2023) Geriatric care management system powered by the IoT and computer vision techniques. *Healthcare* 11. <https://doi.org/10.3390/healthcare11081152>, <https://www.mdpi.com/2227-9032/11/8/1152> (visited on 08/20/2023)
8. Mu H et al (2021) Discovering and analyzing internet of things (IoT) technologies and applications for aged care. *J Phys Conf Ser* 1948(1):012141. <https://doi.org/10.1088/1742-6596/1948/1/012141>, <https://dx.doi.org/10.1088/1742-6596/1948/1/012141>

9. Gharti P (2020) A study of fall detection monitoring system for elderly people through IOT and mobile based application devices in indoor environment. In: 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), pp 1–9. <https://doi.org/10.1109/CITISIA50690.2020.9371773>
10. Nair AS, Antony KC, Sujatha R, (2023) Remote health monitoring of geriatric patients using internet of wearable things. In: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), pp 838–842. <https://doi.org/10.1109/InCACCT57535.2023.10141738>
11. Raptan S, et al (2023) Cardioxy – A novel and portable instrumentation amplifier based and IoT enabled ECG device. In: 2023 2nd International Conference for Innovation in Technology (INOCON), pp 1–5. <https://doi.org/10.1109/INOCON57975.2023.10101028>
12. Karagiannis D, Nikita, KS (2020) Design and development of a 3D printed IoT portable Pillbox for continuous medication adherence. In: 2020 IEEE International Conference on Smart Internet of Things (SmartIoT), pp 352–353. <https://doi.org/10.1109/SmartIoT49966.2020.00066>
13. Cristescu I, Băjenaru L (2021) Elderly smartwatch adoption: a conceptual model development. In: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pp 1–4. <https://doi.org/10.1109/ECAI52376.2021.9515160>
14. Bandyopadhyay S, et al (2022) MediFi : an IoT based wireless and portable system for patient’s healthcare monitoring. In: 2022 Interdisciplinary Research in Technology and Management (IRTM), pp 1–4. <https://doi.org/10.1109/IRTM54583.2022.9791747>
15. Mulfari, D, Minnolo, AL, Puliafito A (2017) Wearable devices and iot as enablers of assistive technologies. In: 2017 10th International Conference on Developments in eSystems Engineering (DeSE), pp 14–19. <https://doi.org/10.1109/DeSE.2017.51>
16. Moatamed B, et al (2016) Low-cost indoor health monitoring system. In: 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 159–164. <https://doi.org/10.1109/BSN.2016.7516252>
17. Crespo-Salgado JJ, et al (2014) Guía básica de detección del sedentarismo y recomendaciones de actividad física en atención primaria. www.clinicalkey.es. <https://www.sciencedirect.com/science/article/pii/S0212656714002911?via%3DIihub> (visited on 11/10/2023)
18. Colomer F, Puig-Ribera A (2022) Inactividad física y sedentarismo. In: AMF 18, pp 66–75. https://salutpublica.gencat.cat/web/.content/minisite/aspcat/promocio_salut/suma_salut/noticies/activitat_fisica/02_Los-principales-problemas-de-salud-Febrero.22.pdf (visited on 11/14/2023)
19. Miguel-Ruiz J, (2022) Medición del gasto energético en adultos utilizando sensores vestibles. spa. MA thesis. Cuernavaca, México.: Tecnológico Nacional de México. <https://rinacional.tecnm.mx/jspui/handle/TecNM/4158> (visited on 03/09/2025)



A Robotic System Integrating the Semmes-Weinstein Monofilament (SWM10g) for Automated Plantar Sensory Evaluation: Novel Approach for Early DPN Detection

César Augusto Moreno Baracaldo^(✉) and Martha Zequera Díaz

Department of Electronics, Pontificia Universidad Javeriana, Bioengineering Lab Baspi-Footlab,
Bogotá D.C, Colombia

{caugustomorenob,mzequera}@javeriana.edu.co

Abstract. The present study introduces an automated device for the controlled application of the Semmes-Weinstein 10g monofilament (SWM-10g), designed for quantitative assessment of plantar sensitivity and early detection of diabetic peripheral neuropathy (DPN). As a common complication of diabetes mellitus, DPN increases the risk of lower extremity ulceration and amputation. Our research introduces an innovative automated application technique for SWM10g, characterized by its unique design, efficient methodology, and significant advantages in reduced size and enhanced portability compared to previous approaches documented in scientific literature [19–21].

Ten healthy Colombian volunteers (4 men, 6 women; aged 20–50 years) participated with University Ethics Committee approval. Assessments covered six regions on each foot's plantar surface, with three repetitions per region, totaling 360 measurements. Results were consistent between both methods, reporting normal tactile plantar sensitivity thresholds below 5.0 g-force with a mean of approximately 1.0 g-force. Participants reported positive sensitivity with the traditional manual SWM-10g monofilament test, which demonstrated 100% specificity. Sensitivity comparisons across all contact points revealed that the 1st metatarsal and calcaneal branch points exhibited significantly lower sensitivity compared to other evaluated points ($p < 0.01$). Gender-based comparative analysis yielded no statistically significant differences.

This research is preliminary and focuses on the innovation of a device for the evaluation and measurement of plantar sensitivity, with clinical evaluation not being its main objective. The clinical assessment and results presented in this study constitute complementary work carried out to test the equipment's functionality, which also explains the limited number of subjects included in the study. Further refinement of the development is recommended before subsequent evaluation in patients with DPN, complemented by comparisons with nerve conduction studies (NCS) and autonomic function testing (AFT) for early detection of sensitivity loss. The methodology establishes groundwork for the quantitative assessment of plantar sensitivity thresholds and demonstrates potential for enhanced accuracy.

Keywords: Diabetes mellitus · Foot ulcer prevention · Medical robotics · SWM-10g monofilament · Electronic measurement system · Embedded

electronic device · Nerve conduction studies (NCS) · Autonomic function testing (AFT)

1 Introduction

According to the World Health Organization (WHO, 2023), approximately 442 million people worldwide suffer from diabetes (types I and II), with at least 62 million affected in the Americas, around 40% of whom are unaware of their condition. This disease is the sixth leading cause of death in the Americas and was responsible for more than 284,000 deaths in 2019 (PAHO, 2022). In December 2021, the International Diabetes Federation (IDF) in its 10th edition of the Global Diabetes Map published that the prevalence of diabetes increases with age, reaching 24%, with a statistical peak mostly reflected in people over 75 years old. It was estimated that by 2024, medical costs will increase to approximately 1,054 billion dollars, with an estimated 783 million diabetic patients worldwide [1]. In Colombia, the prevalence of type 2 diabetes (DM2) ranges from 4% to 8%, depending on the age group. It is noteworthy that there is a high prevalence of overweight (more than 30%) and metabolic syndrome (between 20% and 35% of the population) [2]. The development of DM2 is associated with high levels of BMI, TyG, glucose, and triglycerides [3], and MHO is associated with the increase of DM2 compared to MHNW [4]. The conditions of patients suffering from type 2 diabetes mellitus are diverse, with complications of different kinds, including: diabetic neuropathy, diabetic foot, gait abnormalities, diabetic foot ulcers, diabetic retinopathy, and cardiac abnormalities, pain, muscle numbness, among others. In the U.S., [5, 6].

People with type 2 diabetes mellitus are at risk of foot ulceration due to the increased process of diabetic sensory polyneuropathy (DSPN), which manifests as a decrease or loss of the ability to detect tactile sensations, primarily in the soles of the feet [7].

DSPN and diabetic peripheral neuropathy (DNP) are the most common forms of neuropathy complications (NP). NP is the degenerative condition of the sensory nervous system. DNP presents as damage at the nervous system level and is a consequence of peripheral vascular disease (PVD), which is a microvascular impairment or blockage of blood perfusion where oxygenation is reduced, causing hypoxia of the skin tissues of the foot and nervous tissues [8]. This results in the loss of the terminal nerve and, consequently, the loss of sensory afferents in patients with prolonged and poorly managed diabetes who did not have an early alarm or diagnosis of NPD.

The prevention of diabetic foot complications, specifically diabetic foot ulcer (DFU), requires an approach from the perspective of early warning. This allows for the identification of whether the diabetic patient presents initial symptoms of NPD or decreased plantar sensitivity that predispose them to ulceration and may lead to amputations. Current technologies for the evaluation of diabetic foot primarily focus on measuring temperature and pressure [9], with limited options for assessing sensory perception levels in lower extremities.

The SWM monofilament is the most commonly used method [10, 11] due to its low cost [12], although it has limitations for the early detection of NPD. There are more precise but costly methods, such as nerve conduction velocity (NCS) [13], autonomic

function tests (AFT), and biothesiometers, which are used exclusively in clinics or laboratories under controlled conditions [14]. However, its manual application introduces significant biases due to variability in the application technique among specialists and the lack of specific pressure control. The detection of NPD using a SWM-10g, requires the evaluation of at least 10 application sites, which is complex and demanding for both the medical specialist and the patient [15]. The American Diabetes Association (2017) established guidelines for the detect and similar methods to diagnose “high risk” of ulceration at 3 sites per foot. However, various studies [16, 17], and similar methods to diagnose “high risk” of ulceration at 3 sites per foot. However, various studies discourage its use as the sole diagnostic method, suggesting that it should be complemented with other clinical techniques to improve its sensitivity and specificity, since the traditional method is performed manually and does not allow for the quantitative recording of pressure values to determine the degree of sensitivity [18]. Previous autonomous robotic developments for SWM-10g application, such as Castellano’s system [19], present limitations in clinical settings due to their cost, size, and complexity of use in a clinical environment. These constraints motivated the development of a new technique supported by a robotic system, which offers enhanced viability in terms of clinical implementation and practical deployment in medical settings. [19], present limitations in clinical settings due to their cost, size, and complexity. These constraints motivated the development of this prototype, which offers enhanced viability in terms of clinical implementation and practical deployment in medical offices.

This study uses an automated tool, a custom-built robotic arm, that allows the application of the SWM -10g monofilament at six point on the plantar surface. The main objective of this study was to compare the levels recorded by thated tool against the semi-quantitative results generated by the traditional application tech.

The goal is also to develop a low-cost solution that improves the evaluative capacity by quantitatively indicating the level of plantar sensitivity perception in response to a contact pressure stimulus in diabetic patients at risk of developing early neuropathy, thus facilitating early detection and timely treatment of this condition.

2 Methodology of Analysis

The automation for the tactile sensory evaluation required the design and construction of an electromechanical robotic arm prototype that handles and applies the 10g-SWM in a controlled manner.

The methodology first describes the characterization of the main force sensor and five SWM-10g, followed by the implemented testing protocols for the evaluation of tactile sensory evaluation in a group of normal subjects.

A comparative pilot study was conducted between both monofilament application techniques: The traditional manual technique and the robotic technique implemented. The assessment consisted of determining the participants’ sensitivity perception through two sequential methods: first, the manual application of the (SWM-10g), followed by the robotic application of the same monofilament.

The volunteers were clinically healthy individuals, so no false positives or false negatives were expected in the tests. The comparison was made between semi-quantitative and

quantitative responses, seeking to identify significant differences between the different evaluated points of the foot based on the data recorded with the prototype. Additionally, the study with diabetic individuals is proposed for future research. The analyses were conducted using the statistical program R-Studio.

For the robotic arm development, validation of the load cell as a force sensor for this specific application was required, necessitating a detailed analysis of its measurement characteristics. Initially, the analysis of the force sensor (load cell) was carried out, evaluating its linearity and hysteresis response, as well as physical and mechanical characterization of five Semmes-Weinstein monofilaments (SWM-10g) from a single manufacturer, performing the corresponding statistical analyses for each case.

2.1 Analysis of Physical Characteristics of the Load

The load cell was evaluated through linear regression analysis. The statistical results showed an extremely small p-value ($<2.2e-16$), lower than the established significance level ($\alpha = 0.05$), which allowed rejecting the null hypothesis. The coefficient of determination (R^2) obtained was 0.9982, indicating that the model explains 99.82% of the variability of the data. The slope was statistically significant, confirming the linear relationship between the variables.

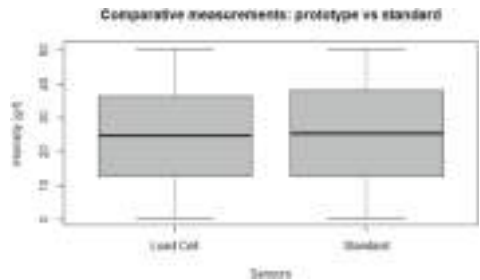


Fig. 1. Box plot, it is observed that the force sensor and the scale coincide in the mean response and variance of their data.

A comparative analysis was conducted with a commercial scale using box plots, Fig. 1. The results confirmed that the load cell presents adequate linearity characteristics in the required operating range (0–40 gf), demonstrating its suitability as a force sensor for this application.

2.2 Characterization of the SWM-10g

The characterization of five commercial SWM-10g monofilaments was carried out, evaluating their linearity in the lift curve before critical buckling, diameter, length, and hysteresis in response to compression and decompression.

1. Linearity

A linear regression analysis evaluated the response in the ascending trajectory before critical buckling force. Results showed an intercept (β_0) of 0.09656 and slope (β_1) of 68.10270, demonstrating a strong linear relationship between advance and measured force ($p < 2e-16$). The coefficient of determination (R^2) was 0.9852, explaining 98.52% of data variability, Fig. 2. Residuals ranged from -1.3840 to 2.5457 , with symmetric distribution around zero, validating the linear model's adequacy and the measurement system's accuracy in the studied range. The 5 monofilaments had a linear and predictable response within their operating range.

2. Repeatability, Sensitivity, and Accuracy

Repeatability was evaluated through five consecutive tests for each monofilament. The results showed consistent measurements with no significant variations between them. However, it was observed that the SWM -10g monofilament experiences a degradation of its elastic properties after multiple uses. The tests revealed that after 10 applications, the monofilament begins to lose its original characteristics. An extensive analysis with more than 30 cycles of use on the same monofilament demonstrated an increase in the critical point up to 8 g-force from its initial value, evidencing the importance of limiting the number of applications per monofilament.

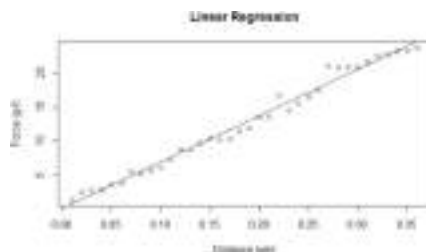


Fig. 2. Linear regression, Linear working section of the SWM-10g under test in the first part of the curve before the critical buckling force point.

Regarding accuracy, the measurements taken in five consecutive repetitions showed a deviation of less than 5% from the nominal value, indicating an acceptable level of reliability in the measurements within the recommended usage range.

3. Precision

In the evaluation of 5 repetitions, the measurements showed values with a precision greater than 95%.

4. Shape and Size

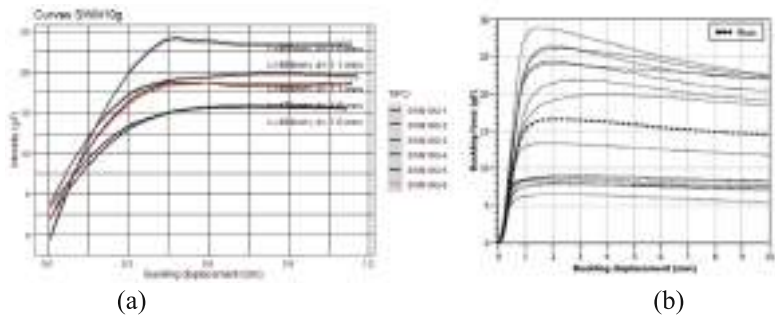


Fig. 3. Behavior curves of five (5) SWM-10g monofilaments are observed (a). The different behavior curves that were performed in this study when the monofilaments are subjected to variable compression coincide with the study of (Castro-Martins, Pinto-Coelho, & Campilho, 2024). Image (b).

Length: between 40.00 mm and 42.00 mm, Diameter: 0.80mm, Material: polyamide #6,6.

5. Comparative Analysis Between SWM-10g Monofilaments

A comparative analysis of five SWM-10g monofilaments using the Wilcoxon test for paired samples revealed statistically significant differences ($p < 0.05$). This finding is demonstrated in Figs. 3 and 4, showing behavior curves and comparative box plots, respectively. None of the evaluated commercial monofilaments met the manufacturer's 10 g-force buckling point specification, all showing higher values. Environmental factors (humidity, temperature) and repetitive use significantly affected the material's elastic characteristics, confirmed through a 30-cycle study of SWM10g-1 monofilament (Fig. 3(a), red curve). These findings align with reported changes in nylon 6.6 behavior under variable conditions.

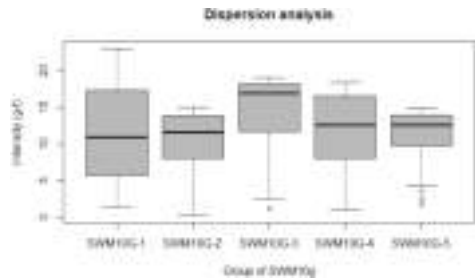


Fig. 4. Statistical comparisons, Box plot as a comparison of the mean differences between the five (5) evaluated SWM-10g. The difference between means and data variability is quite significant.

6. Hysteresis

The hysteresis behavior of the monofilaments was analyzed by recording load and unload curves in the range of 0 to 40 g-force. The Wilcoxon test revealed that there are no statistically significant differences between the compression curves ($p = 0.9037$) and decompression curves ($p = 0.7222$), with a significance level of $\alpha = 0.01$.

These results indicate a consistent and reproducible behavior in the load and unload cycles of the evaluated monofilaments.

Implications of the Analysis

The significant variability found among the evaluated monofilaments highlights the need to implement rigorous protocols for individual calibration and quality control. This heterogeneity in mechanical characteristics is particularly relevant considering that, of the analyzed commercial SWM-10g set, only two monofilaments were close to the nominal buckling strength specification (10 g-force), while the other three presented much higher values.

The results obtained also highlight the influence of environmental factors such as humidity and temperature on the elastic properties of nylon. This phenomenon was confirmed through the analysis of the monofilament SWM10g-1 after 30 cycles of use and exposure to variations in ambient temperature Fig. 3 (a), corroborating the significant changes in the mechanical properties of nylon 6.6 documented in the literature.

3 Development of an Automated Evaluation Tool: Design and Construction of a Robotic Arm

Existing methods for automated plantar sensitivity perception assessment are highly limited and remain in developmental stages. We identified a device developed by [19], which uses the SWM10g in a very accurate and comprehensive electromechanical construction for evaluation, but it has limitations such as its size and weight. Other automated methods, such as the one by [20], present a design focused on automating two (2) SWM10g units for sensitivity assessment using the aesthesiometry method, but it remains a manually applied instrument. Other methods utilize computerized technology with electrical or pressure stimuli to measure evoked responses as explained in the work of [21], but these systems are very expensive and require specialized spaces for their application.

Our development is a device that positions the SWM10g on an easily operated adjustable arm, offering advantages over the aforementioned alternatives, such as transport convenience, adaptability to any table, desk, or medical stretcher, and flexibility to change parameters like application speed, angles, and position. Figure 5 shows the block diagram of the developed equipment and Fig. 6 shows the actual prototype with its main parts. As seen in the Fig. 6, this prototype uses the SWM-10g monofilament at the advancing tip, which is mounted on a load cell that is in turn installed on a sliding base system. In Table 1, a comparison is shown between our design and existing devices.

This system is driven by a stepper motor (SANYO SANMOTION2-Phase Stepping Motor F2 SS2281-5731) with the ability to move forward and backward at different speeds, controlled by the TB6600 driver and an Arduino MEGA microcontroller.

The system advances with the monofilament at a speed of 0.3 mm per second to reach the plantar surface, acting with a gentle touch of compression on this surface,

Table 1. Technical Functionality Comparison Between Devices.

Author	Strengths	Weaknesses
Castellano et al. [19]	Vertical tables automation with 3D capabilities, AI-powered foot image analysis, and robust construction	Oversized, not suitable for use in medical offices, cumbersome to operate
Ganeshmurty et al. [20]	Utilization of two SWM10g unnnits to improve distance precision between them	Still requires manual application to the selected plantar points
Behforootan et al. [21]	Use of specialized equipment for sensitivity evaluation through evoked pressure stimuli. Device with high sensitivity and specificity in testing	Equipment that must be used in appropriate spaces
Robot Arm	Easy handling and operation. Can be easily transported with simple installation, making it highly suitable for use in medical offices. Control of SWM10g application speed. Eliminates bias from the SWM10g material	Equipment that lacks complete positioning automation with more degrees of freedom. Does not incorporate AI for autonomous self- positioning. Still has weight and size that could be reduced

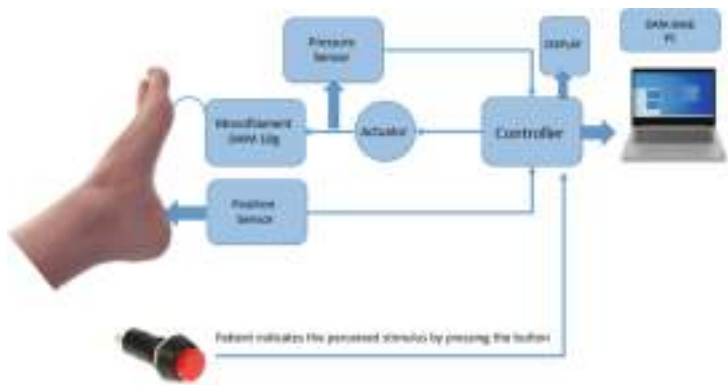


Fig. 5. Block schematic of the prototype. Process for recording measurement and constitutive blocks of the developed equipment and right - Algorithm, showing the action line of the sequence of the test performed by the prototype.

which allows recording the force exerted on the load cell (mini double shear load cell – 500 g, with 900-Ω impedance). The volunteer indicates the moment they perceive contact with the monofilament by pressing a switch button (Fig. 7 b), which is connected to the mentioned microcontroller; this button acts as feedback between the subject and the device, allowing the recording of the force exerted by the monofilament on the contact

point. This reading is displayed on the prototype’s display (1.9” OLED display) and is stored in the ROM memory of the microcontroller. This record corresponds to the force in grams that the subject perceived during the test. The load cell is connected to a signal amplifier, and this output signal is converted to digital format with 24-bit resolution using the HX711 chip with 10 Hz sampling.

The prototype has a time-of-flight distance sensor that allows verifying if the test area is within the operational range of the equipment, displaying on the screen when the contact area is outside the front advance limits of the equipment.

The prototype assembly includes a control box (Fig. 8) that allows operating the equipment through three control buttons: test start, sensor tare, and emergency stop. The entire assembly is mounted on a movable scissor arm that allows movement across the plantar surface to position the tip of the monofilament over the target point.

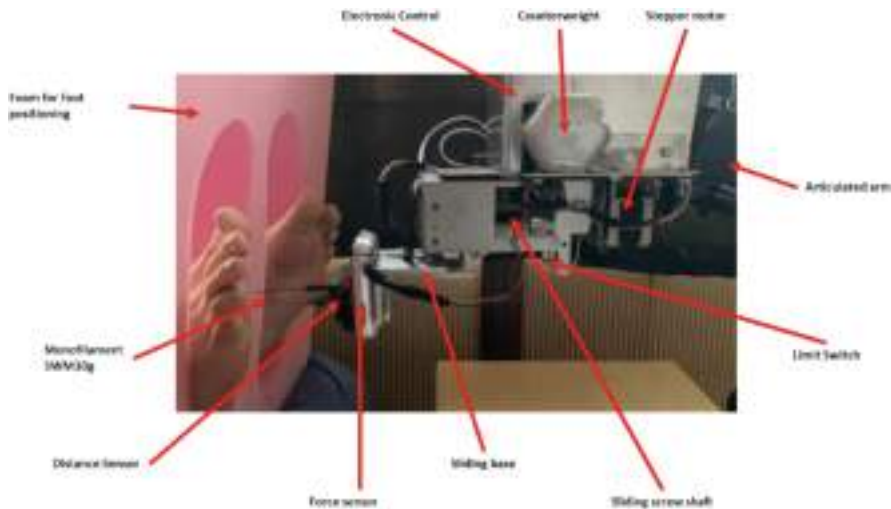


Fig. 6. Developed prototype – In the patent process. The electronic and mechanical parts that make up the equipment are named.

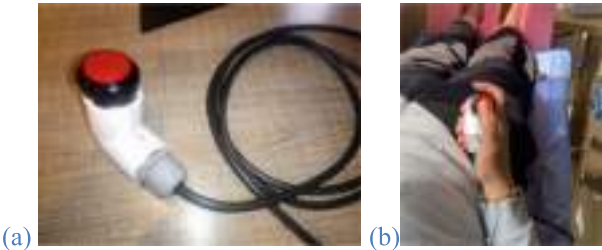


Fig. 7. (a). Feedback button, which through its manual operation indicates the moment when the volunteer perceived the contact of the monofilament. (b) Patient pressing the button when they perceive the touch of SWM10g on their foot.

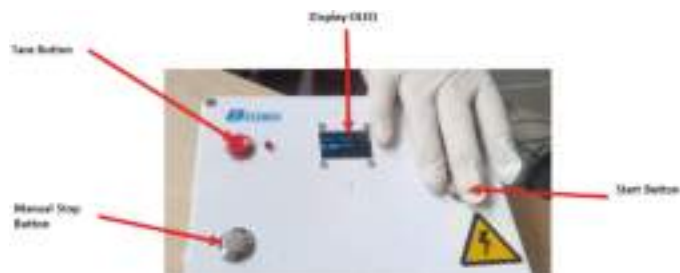


Fig. 8. Control box that includes the microcontroller system and control buttons.

4 Manual and Robotic Monofilament Application Assessment

4.1 Population and Exclusion Criteria

The study was conducted in Bogotá, Colombia, under the direction of Dr. Gustavo Castro, a neurology specialist, and with approval from the Ethics and Research Committee (CIE) of the Pontificia Universidad Javeriana (registration FID-24-462). Ten Colombian participants (4 men, 6 women) aged between 20 and 50 years were randomly selected, who provided their written informed consent.

The inclusion criteria included: clinically normal health status, absence of motor impairments, chronic illnesses, or diabetes diagnosis. The initial medical evaluation, conducted by the specialist, included the recording of the Body Mass Index (BMI). The study group had an average BMI of 25.58 kg/m^2 , slightly above the normal range established by the World Health Organization ($18.5\text{--}24.9 \text{ kg/m}^2$). No participant showed health alterations that could interfere with the study results (Table 2).

4.2 Application of the SWM-10g Monofilament Using the Traditional Manual Technique

The evaluation of plantar sensitivity was performed using the traditional manual technique of the SWM-10g monofilament.

Participants were positioned supine on an examination table, with bare feet in a relaxed state. Six specific anatomical points on each foot were evaluated (12 total locations, Fig. 9 a - c). 1. Distal first phalanx (Hallux), 2. Distal first metatarsal, 3. Distal fifth metatarsal, 4. Saphenous nerve branching (mid-plantar area), 5. Lateral plantar branching (mid-plantar area), 6. Calcaneal branching.

The specialist applied the monofilament to its bending point at each location, following an ascending order from the Hallux. Each point received three consecutive evaluations. Participants verbally reported the perception of the tactile stimulus, and the responses were recorded in an evaluation log.

Table 2. Age, weight, and height data of the participants. Categorized BMI - 18.5: Underweight, 18.5–24.9: Normal, 25.0–29.9: Overweight, 30.0 or more: Obesity. World Health Organization Bulletin. “Body mass index - BMI. “WHO Technical Report Series.

Gender	Age	Weight (kg)	Stature (m)	BMI	Average BMI
F	46	69	1.43	33.74	24.95
F	42	62	1.6	24.22	
F	46	68	1.57	27.59	
F	35	51	1.64	18.96	
F	47	65	1.61	25.08	
F	25	49	1.56	20.13	
M	27	72	1.65	26.45	26.51
M	49	80	1.71	27.36	
M	22	74	1.72	25.01	
M	23	75	1.66	27.22	
			Average	25.58	

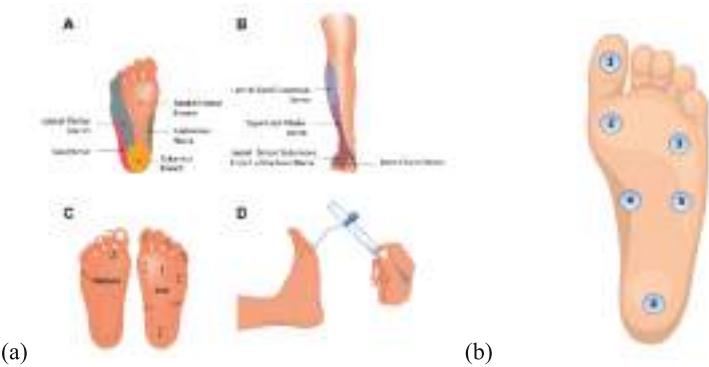


Fig. 9. Evaluation sites of the plantar surface. (a). Different evaluation points. Taken from [17], Right. Six (6) sites per foot were chosen, which were evaluated using the SWM-10g.- Own elaboration. (b).

4.3 Application of Semmes-Weinstein 10g Monofilament Using Robotic Arm Prototype

For the test, the subject was positioned in a supine position on the stretcher, with legs extended and in a state of relaxation. A perforated EVA foam was placed on the feet to prevent lateral bobbing.

When the prototype (Robotic Arm) with the monofilament is positioned frontally towards the contact point to be evaluated, the process begins through the equipment controls, allowing it to advance linearly and frontally towards the surface of the foot

autonomously. If the distance between the tip of the monofilament and the surface is greater than 0.5 cm, the advance is made at a speed of 0.5 cm per second until it is positioned 0.5 cm from the target, where the speed is reduced to 0.3 mm per second. This allows for a soft and light contact of the monofilament on the point of the plantar surface. This process is repeated three times on the same contact point. The advance of the monofilament is programmed not to exceed the critical buckling force of the monofilament (10 gf).

The difference of this proposal compared to other methods lies in that only the linear curve fraction of compression stress of the monofilament is used before reaching its critical buckling point. This avoids manufacturing bias of the monofilaments, as various SWM-10g monofilaments tested in the laboratory do not meet the specifications listed in the manufacturer's data sheet, as seen in Fig. 3. The repeatability in the use of the (10g-SWM) significantly increases through this application technique, which controls the variability inherent to the material's elastic properties. This characteristic allows the replacement of the SWM-10g with a monofilament having a higher critical buckling point, thus expanding the range of available pressures within its characteristic working curve prior to the critical buckling point.

This autonomous device for the application of the SWM-10g eliminates operator-associated biases, such as speed and angle of application. Additionally, it provides a quantitative reading of the perceived force, which can contribute to clinical decision-making.

5 Comparative Analysis Between Manual Technique and that Applied by the Automated Prototype

A comparative analysis was conducted between the threshold sensitivity recorded using the traditional method (dichotomous yes/no response) and the quantitative data obtained through the automated prototype at six evaluation points per foot. Since the data did not meet the normality assumption according to the Shapiro-Wilk test, the non-parametric Wilcoxon test was used for the comparative analyses.

The study evaluated multiple variables, differences in sensitivity between genders, and comparisons between the left and right foot. This multifactorial analysis was conducted with the aim of establishing a reference baseline for future research.

6 Results

A total of 360 measurements were recorded corresponding to the evaluation of ten clinically healthy volunteers (six women and four men), with no visible podiatric alterations or calluses. The protocol included the evaluation of six points per foot, with three contact repetitions per point.

All participants signed an informed consent authorizing data processing and acknowledging the minimal risks of the procedure. Additionally, they declared not to have a diagnosis of chronic or neurological diseases.

The group results showed differentiated patterns of sensitivity according to the anatomical location. Areas with a thicker stratum corneum, such as the heel region

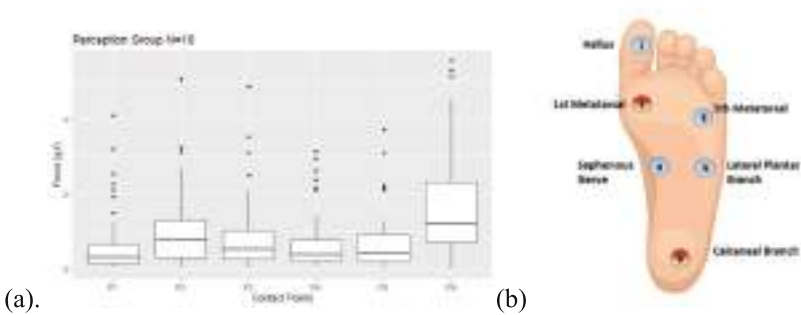


Fig. 10. (a). - Perception intensity graph recorded in the test group. (b). - Plantar distribution of sensitivity to the SWM-10g monofilament. The areas in red gradient indicate lower sensitivity to contact, while the areas in blue gradient represent areas of higher sensitivity.

and the middle plantar branching (P2 and P6), exhibited greater variability in sensitivity thresholds and lower sensitivity. In contrast, the areas less exposed to pressure during walking, corresponding to the saphenous nerve and the base of the big toe (P1, P3, P4, and P5), showed higher sensitivity indices, Fig. 10.

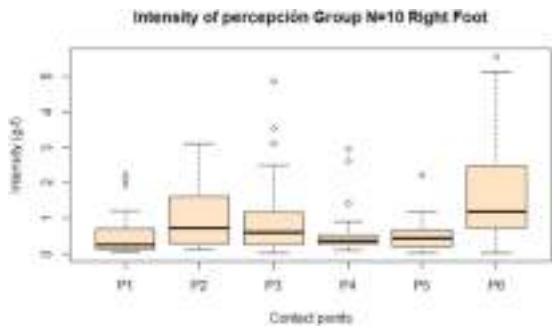


Fig. 11. Graph of Perception Intensity recorded in the test group on the right foot only, points P1 to P6 correspond to the evaluated plantar locations. It is the response of the mean sensitivity perception reported by the participants in the test conducted with the automated method.

Since the data did not follow a normal distribution according to the Shapiro-Wilk test, the Wilcoxon test was used for comparative analysis. Points P1, P3, P4, and P5 showed similar perception thresholds, ranging from 0.3 to 0.5 g-force. On the other hand, points P2 and P6 exhibited significantly higher thresholds, with P2 approaching 1.0 g-force and P6 exceeding this value. This distribution of thresholds is visualized in the box plot Fig. 10 (a) and in the plantar representation Fig. 10 (b).

The areas corresponding to points P1, P3, P4, and P5 exhibited significantly lower sensitivity thresholds compared to P2 and P6, which correlated with areas of greater load during walking and greater development of the stratum corneum. The analysis revealed that, in healthy subjects, regardless of gender, sensitivity thresholds remained below 5.0 g-force in all evaluated areas, with a mean of approximately 1.0 g-force. To

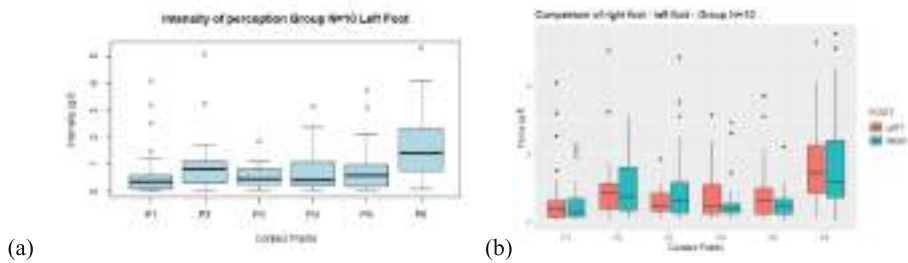


Fig. 12. (a). Graph of Perception Intensity recorded in the test group on the left foot only. (b) comparative of means of sensitivity of both feet, where no statistically significant differences are found by point.

minimize the inherent variability in measurements, a protocol of three repetitions per point was implemented. Data that showed inconsistencies in perception reporting were excluded from the analysis. Two potential sources of methodological bias were identified: anticipation of the stimulus due to the visibility of the equipment and variability in response time when perceiving the stimulus.

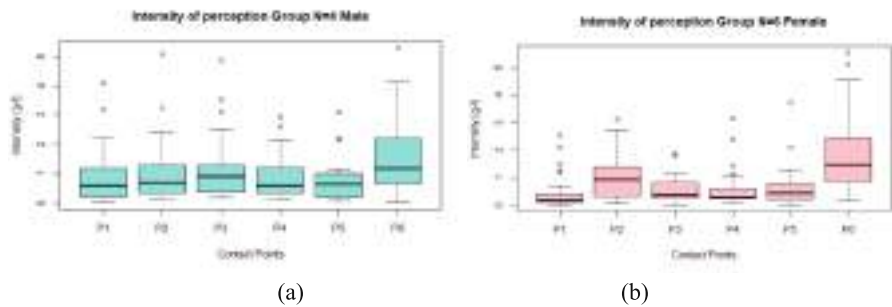


Fig. 13. Comparison of plantar sensitivity perception between male and female groups. (a). Male group showing uniform sensitivity across all plantar points, with higher variability in points 2 and 3. (b). Female group displaying significantly lower sensitivity in points 12 and 6 compared to other plantar regions.

The analysis using box plots did not reveal statistically significant differences in sensitivity patterns when comparing the right and left foot, nor between genders, as can be seen in Figs. 11 and 12, respectively.

The analysis of graphs 11 and 12 reveals slightly less variability in the left foot measurements, though this difference is not statistically significant enough to draw definitive conclusions. When comparing results between genders, as illustrated in Figs. 13 and 14, no significant differences emerged between male and female participants. Specifically, Fig. 13 demonstrates nearly identical mean values and data variability patterns for both gender groups, with particularly similar distributions observed in the right-foot measurements.

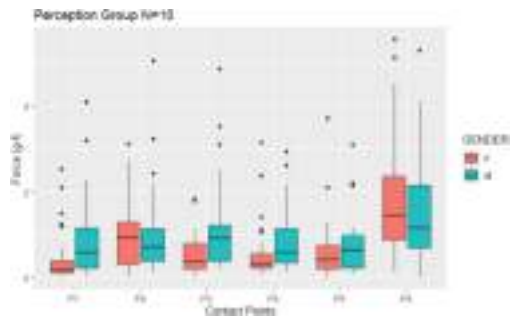


Fig. 14. Graph of Perception Intensity recorded in the group of men.

When performing the traditional application technique of the SWM10g, all volunteers responded that they DID perceive contact with the monofilament. Since it was not possible to demonstrate a range or level of force to compare except for the 10gf bending force reported by the manufacturer, a semi-quantitative graph was created, which is shown in the following figure (Fig. 15).

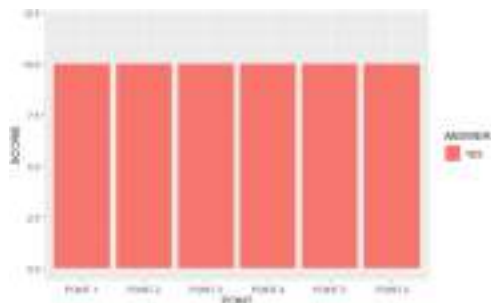


Fig. 15. Data representation when using the traditional application technique of the SWM10g – It is a semi-quantitative technique. All volunteers responded that they did perceive the touch with the monofilament at each of the points evaluated on the sole of the foot.

The sensitivity and specificity analysis of the instrument showed an effectiveness of 100%, with no records of false positives or false negatives. This result was expected since the study population consisted exclusively of clinically healthy individuals, with no history of sensory disorders, diabetes, or other conditions that could affect the results. All participants exhibited perception thresholds below 5 g-force, demonstrating that the developed instrument has greater precision in quantifying sensitivity compared to the traditional application technique.

It is proposed to give a representative color to each level of force ranges of the different SWMs and that graphically they can give a visual indication of the results at each point of the foot (Fig. 16 a and b) that can be helpful and supportive in a medical diagnosis.

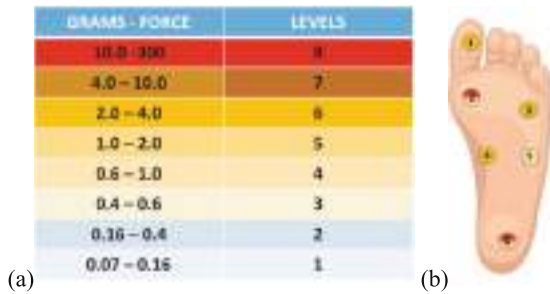


Fig. 16. Pressure intensity reference scales. (a). Commercial Semmes-Weinstein Monofilament (SWM) pressure levels ranging from 0.07g to 300g according to references [17, 20]. (b). Proposed diagnostic intensity scale for early detection of Diabetic Peripheral Neuropathy (DPN).

7 Discussion

This study is based on previous research [15, 17, 20] and the recommendations of the American Diabetes Association regarding the plantar application of the SWM-10g monofilament. The literature agrees in pointing out significant limitations in the current technique, including the complexity of the procedure, the high workload for the specialist, and the need for automation in the evaluation. The automated device developed addresses several critical limitations identified in current assessment methods:

Methodological Standardization: Previous research has identified significant variability in manual application techniques [22], including force exerted, contact speed, and application angle, which compromise diagnostic effectiveness. Various researchers [9, 16, 23, 24] agree that the method is reliable only in cases of established DPN. Our automated approach standardizes these parameters, potentially improving test reliability.

Quantitative Assessment: The device overcomes the traditional method's primary limitation of providing only qualitative (yes/no) responses by delivering precise force measurements [25]. This quantitative approach enables the detection of subtle changes in sensitivity thresholds, particularly valuable for early-stage DPN detection.

Technical Innovation: The study's novel approach of utilizing the linear portion of the monofilament's force-displacement curve before the buckling point represents a significant technical advancement. This method mitigates the variability found in commercial monofilaments' specifications, as demonstrated in our characterization study where significant deviations from manufacturer-specified buckling forces were observed. This approach aligns with the current trend towards automated solutions for the early assessment of diabetic foot, as confirmed by [19].

Clinical Implications: The findings suggest that anatomical location significantly influences sensitivity thresholds, with areas of thicker stratum corneum showing higher thresholds.

Gender and Bilateral Comparisons: The absence of significant differences between genders and between left and right feet in healthy subjects establishes a baseline for future comparative studies with diabetic populations. A lower BMI may be related to a better perception of plantar sensitivity as observed in the group of women whose

average BMI is 24.95 kg/m² showing better sensory response at all evaluated points compared to the sensory response shown in men.

7.1 Limitations

This study presents significant methodological limitations. The small sample of ten clinically healthy volunteers prevents the generalization of results to patients with diabetic peripheral neuropathy (DPN). Additionally, the absence of patients with diagnosed diabetes or neuropathy, as well as the lack of comparison with reference methods such as nerve conduction studies (NCS) or autonomic function tests (AFT), constitute important limitations of the experimental design.

The technical limitations of the prototype include the sensitivity of the distance sensor to ambient lighting, the considerable dimensions of the load cell, and the need for manual positioning in front of the contact point. The overall size of the device affects its portability and clinical applicability.

For future research, it is proposed:

- Expand the study to a larger and more heterogeneous population, including diabetic patients.
- Conduct comparative validation with reference methods (NCS and AFT).
- Develop longitudinal studies to assess the predictive value of the method in the development of neuropathy.
- Optimize the device design to improve its portability and clinical usability, including full automation of positioning.

8 Conclusions

This research demonstrates the feasibility of optimizing the traditional technique of applying the SWM-10g through an automated device that controls multiple variables, reducing the biases inherent in manual application. The developed device achieved a sensitivity and specificity of 100% in clinically healthy subjects, with no reports of false positives or negatives.

The characterization of commercial SWM-10g monofilaments confirmed the findings of [26], revealing significant discrepancies between the manufacturer's specifications and the actual behavior of the instruments. The designed device overcomes this limitation by basing its measurements on the applied pressure rather than on the bending of the monofilament.

The instrument demonstrated the ability to record application forces in a range of up to 300 g-force, allowing for the quantitative assessment of various levels of sensitivity. This feature facilitates the precise evaluation of the threshold of protective sensitivity perception in different populations, including patients with various degrees of diabetic peripheral neuropathy (DPN).

The results in healthy volunteers showed sensitivity thresholds below 5 g-force, with variations according to the anatomical region. The heel areas and the medial plantar branching L4-L5 showed lower thresholds compared to the areas of the saphenous nerve and the lateral plantar branching. The developed device represents a significant advance

in the evaluation of plantar sensitivity, providing a more objective and quantitative method for the early diagnosis of DPN.

Limitations and future work:

- The small sample ($n = 10$) limits the generalization of results
- Need for validation in diabetic population and comparison with reference methods (NCS, AFT)
- Technical optimization required:
- Complete automation of positioning using AI
- Miniaturization of the force sensor
- Improvement in ambient light compensation
- Refinement of the distance measurement system

References

1. Huang H et al (2023) Physical therapy in diabetic foot ulcer: research progress and clinical application. *Int Wound J* 20(8):3417–3434
2. Aschner P (2010) *Diabetología*, pp 95–100
3. Fang-fang Z et al (2024) Association of triglyceride glucose - Body Mass Index & TyG -BMI) and risk of developing type 2 diabetes mellitus. *Mod Prev Med* 51(24):4591–4596
4. Wang X, Wu Y, Wang Y, Zhou J, Liu T (2024) Relationship between metabolically healthy overweight/obesity and risk of type 2 diabetes in different ethnicity: a prospective cohort study in southwest China. *BMC Public Health* 24(1):1–11
5. Nie Q et al (2024) Analysis of sleep for the American population: result from NHANES database. *J Affect Disord* 347(2023):134–143
6. Mojaver A et al (2017) The effect of footwear adapted with a multi-curved rocker sole in conjunction with knee-ankle-foot orthoses on walking in poliomyelitis subjects: a pilot study. *Disabil Rehabil Assist Technol* 12(7):747–751
7. Betancourt-Sandoval JA, Covarrubias-Gómez A, Guevara-López U, (2008) Mecanismos generadores y consideraciones terapéuticas futuras de la polineuropatía diabética dolorosa. *Rev Mex Anestesiología* 31(1):28–36
8. Kamei N et al (2005) Effectiveness of semmes-weinstein monofilament examination for diabetic peripheral neuropathy screening. *J Diabetes Complications* 19(1):47–53
9. Khandakar A et al (2024) Laser-induced graphene-based smart insole to measure plantar temperature. *IEEE Sens J* 24(2):1190–1199
10. Yoo M et al Pilot study of exercise therapy on painful diabetic peripheral neuropathy *Pain Med. (United States)* 16(8):1482–1489
11. Richard JL, Reihes L, Buvry S, Goletto M, Faillie JL (2014) Screening patients at risk for diabetic foot ulceration: a comparison between measurement of vibration perception threshold and 10-g monofilament test. *Int. Wound J* 11(2):147–151
12. Rodríguez-Sánchez B, Peña-Longobardo LM, Sinclair AJ (2020) Cost-effectiveness analysis of the Neuropad device as a screening tool for early diabetic peripheral neuropathy. *Eur J Heal Econ* 21(3):335–349
13. Bhor D et al (2023) Responsive transcutaneous electrical stimulation for management of diabetic foot neuropathy. In: 2023 IEEE 8th International Conference (I2CT), Convergence in Technology (I2CT), 2023 IEEE 8th International Conference for, pp 1–6
14. McGill M, Molyneaux L, Spencer R, Heng LF, Yue DK (1999) Possible sources of discrepancies in the use of the Semmes-Weinstein monofilament: Impact on prevalence of insensate foot and workload requirements. *Diabetes Care* 22(4):598–602

15. Zhang Q et al (2018) Easier operation and similar power of 10g monofilament test for screening diabetic peripheral neuropathy. *J Int Med Res* 46(8): 3278–3284
16. Aithal V, Bhat S (2024) Semmes Weinstein monofilament test for detection of diabetic peripheral neuropathy: sensitivity and specificity. *Egypt J Neurol Psychiatry Neurosurg* 60:1
17. Frade MAC et al (2022) Evaluation of altered patterns of tactile sensation in the diagnosis and monitoring of leprosy using the Semmes-Weinstein monofilaments. *PLoS One* 17:1–20
18. Meling S et al (2023) The composite autonomic symptom score 31 questionnaire: a sensitive test to detect risk for autonomic neuropathy. *J Diabetes Res* 2023:4441115
19. Castellano VK et al (2023) Plantar threshold sensitivity assessment using an automated tool—Clinical assessment comparison between a control population without type 2 diabetes mellitus, and populations with type 2 diabetes mellitus, with and without neuropathy symptoms. *PLoS One* 18:1–22
20. Ganeshmurthy MS, Periyasamy R: Design of a standardised semi-automated sensory device for two point discrimination testing in the foot sole. In: 2020 Sixth International Conference on Bio Signals, Images, and Instrumentation (ICBSII), pp 1–6 (2020)
21. Behforootan S, Chatzistergos PE, Chockalingam N, Healy A, Naemi R (2019) Localized pressure stimulation using turf-like structures can improve skin perfusion in the foot. *Microcirculation* 26(6): 1–11
22. Feng Y, Schlösser FJ, Sumpio BE (2011) The Semmes Weinstein monofilament examination is a significant predictor of the risk of foot ulceration and amputation in patients with diabetes mellitus. *J Vasc Surg* 53(1): 220–226.e5
23. Dros J, Wewerinke A, Bindels PJ, Van Weert HC (2009) Accuracy of monofilament testing to diagnose peripheral neuropathy: a systematic review. *Ann Fam Med* 7(6):555–558
24. Shrestha S, Gorhaly MP, Bajracharya MR (2021) Diagnostic accuracy of monofilament test to detect diabetic neuropathy. *J Adv Intern Med* 10(1): 20–25
25. Dunker Ø et al (2023) Diagnostic accuracy of the 5.07 monofilament test for diabetes polyneuropathy: influence of age, sex, neuropathic pain and neuropathy severity. *BMJ Open Diabetes Res. Care* 11(6): 1–9
26. Castro-Martins P, Pinto-Coelho L, Campilho RDSG (2024) Calibration and Modeling of the Semmes–Weinstein Monofilament for Diabetic Foot Management. *Bioengineering* 11(9):886



Enhancing Cognitive and Motor Skills Through Personalized Digital Games

Petr Novak¹  and Lenka Lhotska^{1,2}  

¹ Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Jugoslavských Partyzanu 1580/3, 160 00 Prague 6, Czech Republic

lhotska@cvut.cz

² Faculty of Biomedical Engineering, Czech Technical University in Prague, nam. Sitna 3105, 272 01 Kladno, Czech Republic

Abstract. Maintaining cognitive function and fine motor skills is crucial for healthy aging, enabling older adults to remain independent, engaged, and enjoy a high quality of life. However, both cognitive abilities and motor skills can decline with age, impacting daily activities and overall well-being. Cognitive games, particularly those incorporating a motor component, offer a promising approach to address this challenge. In the paper we describe our ongoing research that is focused on design and development of a suite of digital games and tasks designed to train fine hand motor skills, eye-hand coordination, perception, visual field, and other cognitive abilities. These games are specifically tailored for individuals who cannot use standard mouse and keyboard interfaces, primarily targeting the elderly, post-stroke patients, and individuals recovering from upper limb fractures. We present already developed games and results of analysis based on user acceptance testing. In the conclusion we propose future enhancements.

Keywords: cognitive game · motor skills · cognitive impairment · personalization · user acceptance

1 Introduction

Maintaining cognitive function and fine motor skills is crucial for healthy aging, enabling older adults to remain independent, engaged, and enjoy a high quality of life. However, both cognitive abilities and motor skills can decline with age, impacting daily activities and overall well-being. Cognitive games, particularly those incorporating a motor component, offer a promising approach to address this challenge. These games provide a stimulating and engaging way to exercise both cognitive functions, such as memory, attention, and problem-solving, and fine motor skills, such as hand-eye coordination, dexterity, and precision. By combining mental and physical activity, these games can offer a more holistic and effective approach to maintaining and even improving cognitive and motor function in older adults. The interactive and often personalized nature of these games can also increase motivation and adherence, making them a valuable tool for promoting healthy aging and supporting individuals in maintaining their independence and well-being. This document outlines the development and enhancement of a suite of such games, designed specifically for this purpose.

This paper aims to address the critical need for cognitive games that actively engage fine motor skills, particularly beneficial for aging populations. While existing digital cognitive games often prioritize mental activity via tablets, mobile phones, or virtual reality, they frequently neglect the crucial role of hand and eye-hand coordination. Recognizing this gap, our ongoing research focuses on the design and development of a suite of digital games and tasks. These are specifically tailored to train fine hand motor skills, eye-hand coordination, perception, visual field, and other cognitive abilities, targeting individuals who cannot use standard mouse and keyboard interfaces, such as the elderly, post-stroke patients, and those recovering from upper limb fractures. To ensure the games' relevance and appeal, we have actively involved seniors in the design process, incorporating their preferences regarding themes and physical objects. The subsequent sections of this paper will provide a brief overview of related work, followed by a detailed description of the developed games and the user-defined requirements that informed their design. Finally, we will discuss the results achieved, and propose future developments and enhancements to this suite of games.

2 Related Work

Research in cognitive games has been developed in recent decades rapidly and currently we can identify several key directions. A crucial area is embodied cognition, investigating how physical interaction and movement influence cognitive processes like learning, remembering, and problem-solving. Relatedly, research on tangible user interfaces (TUIs) can inform the design of games blending physical and digital elements [1]. This research examines the impact of providing tangible service objects to consumers. Specifically, it explores the conditions under which these objects influence consumers' perceptions and behaviors. The study focuses on how a physical connection with a tangible object can foster a stronger psychological bond with the associated service. Ultimately, the paper investigates whether this enhanced psychological connection translates into increased behavioral intentions, such as repeat purchases or positive recommendations. This principle can be utilized in design of cognitive games with tangible objects that contribute to enhancement of cognitive abilities. A good example is spatial learning task [2]. The potential of mixed reality (MR) and augmented reality (AR) to create interactive games incorporating physical objects and movement are continuously investigated. There are already many various applications in different fields, one of them being healthcare [3]. In recent years, several studies were performed with seniors where virtual reality was used as a means for keeping seniors active [4] and/or combining physical exercise with gaming [5]. Studies on serious games, designed for purposes beyond entertainment, are relevant, particularly those using physical interaction for cognitive rehabilitation, assessment, or training [6]. Advancements in sensor technology and movement tracking, including wearable sensors and computer vision, offer opportunities to capture and analyze user interactions. Human-computer interaction (HCI) research contributes by examining the usability and effectiveness of physical-digital games for cognitive purposes. Specific research directions include utilizing tangible object manipulation and movement tracking for cognitive rehabilitation after stroke. Developing interactive games with tangible objects and familiar activities for dementia care is a significant

research direction. Research activities also focus on utilization of the above mentioned technologies allowing measurement and offering feedback for users when performing physical exercise with the aim to prevent and postpone loss of muscles, cognition and functional performance [7]. The reviewed literature reveals a predominant focus on cognitive games that primarily target mental activity, often utilizing tablets, smartphones, or virtual reality. These games frequently overlook the importance of incorporating fine motor skills and eye-hand coordination, particularly crucial for aging populations and individuals with specific motor limitations. For example, Lee and Yii [8] and Skikos and Goumopoulos [9] present games designed for tablets that primarily focus on cognitive functions like memory, attention, and language, with limited emphasis on fine motor skills. Similarly, Choi and Lee [10] explore the potential of adapting an existing computer game for cognitive enhancement in seniors, but the focus remains on cognitive benefits derived from gameplay that relies on traditional input methods. While Guzmán et al. [11] review virtual reality games for cognitive rehabilitation, they highlight the lack of applications that simultaneously address multiple cognitive domains and incorporate dynamic difficulty adjustment.

Although some studies, such as Kleschnitzki et al. [12] and Skikos et al. [13], incorporate physical interaction to some extent, they do not explicitly focus on fine motor skill training. Furthermore, systematic reviews and meta-analyses by Abd-alrazaq et al. [14] and Zuo et al. [15] confirm the effectiveness of serious games for improving cognitive abilities in older adults, but they do not specifically address the role of fine motor skills and eye-hand coordination in cognitive interventions.

This gap in the literature highlights the need for cognitive games that actively engage fine motor skills and eye-hand coordination, particularly for individuals who may have difficulties using traditional input devices. Our research aims to address this need by developing and evaluating a suite of digital games that specifically target these skills, with a focus on accessibility and user-centered design. One of the challenges is finding a good balance between complexity of applied technologies, demands on the personnel skills, financial demands, and the effect on health state of aging population.

3 Digital Games with Objects

3.1 Current State

For many years, our team has been deeply engaged in developing digital games that incorporate physical objects, recognizing the powerful synergy between cognitive and motor stimulation. Collaborating closely with geriatrists and gerontologists, we understand the significant positive impact of combined physical and cognitive activities on the health and well-being of older adults. Our goal has been to design and develop games that are not only effective but also accessible and user-friendly. We prioritize ease of installation and play, ensuring that the games are simple to learn and enjoy, even for those less familiar with technology. Furthermore, we strive to keep the costs low to maximize accessibility. Inspired by the social dynamics of traditional board games, we aim to create experiences that can be enjoyed by multiple players simultaneously, fostering interaction and engagement. Beyond these core principles, we also focus on creating games that are adaptable to individual needs and abilities, offer personalized

feedback, and can be easily integrated into various care settings. In this process, we used the co-creation approach and collaborated closely with seniors in the Gerontology Centre Prague. The volunteers (3 males and 2 females, 75+) gave us a valuable feedback on design, control, graphical layout, and themes of the games. The following text details the current state of development for these object-based digital games. The existing suite comprises four main types of tasks/games:

- **Games with NFC (GamesNFC):** These games utilize a mobile phone with an NFC (near field communication) reader and NFC tags placed on a physical game board or separate cards. The mobile app guides the user to locate, count, or remember sequences by interacting with the NFC tags. The game board and tasks can be customized with various symbols, characters, numbers, or images. Examples are shown in Figs. 1 and 2.



Fig. 1. An example of a game using a mobile phone with a NFC reader and NFC tags placed on cards.



Fig. 2. Another example of a game using a mobile phone with a NFC reader and NFC tags placed on cards.

- **Games with Dice (GamesCube):** These games employ a Bluetooth-enabled playing cube with symbols or characters on its faces. The mobile app prompts the player to rotate the cube to display a specific side. Tasks involve finding or remembering sequences, and the cube faces can be customized. An example is shown in Fig. 3.
- **Videos + Pictures + Questions + Answers (VidPicAskAns):** This application facilitates the creation of slide-show style tasks, including presentations, tutorials, educational courses, puzzles, quizzes, and reminiscence therapy. Images and videos can be incorporated, and audio accompaniment can be added. The app allows for interactive questions and answers, with automatic evaluation, and can be controlled via touchscreen, mouse, keyboard, or specialized buttons (see Fig. 4).



Fig. 3. An example of a game with the cube.

- **Attention + Reaction (ObserveReact):** This application is designed to train attention, reaction time, and visual field. It uses large, mechanically resistant buttons that provide light and sound stimuli, requiring the user to press a button in response. The tasks offer automatic evaluation (see Figs. 5 and 6).

All games share the following characteristics: they do not require a mouse or keyboard, rely on graphical and voice output, allow for custom task creation, support downloadable content, and enable progress saving. Basic manuals and a simple website for downloading content are currently available. The games have been tested with target users (volunteers, who helped in the design process), and initial feedback has been incorporated into the design.

The designed games were tested with potential users (seniors) repeatedly. We discussed with them the layout, graphical design of the games, suitability of control elements, understanding of the game principle. The next step were interviews and testing with professional staff at care homes. It appeared that they were open to the option of



Fig. 4. An example of a quiz.



Fig. 5. An example of a game for training attention and reaction time.

creating their own games and exercises. They also suggested some improvements and adjustments of the objects in the games. Based on user testing and feedback, a range of enhancements were requested to improve the games' usability, accessibility, and therapeutic value. Crucially, the designed games also facilitate the collection of detailed gameplay data, including reaction times and accuracy. This data will be used for quantitative performance analysis, enabling comparisons to established neuropsychological assessments such as the CANTAB (Cambridge Neuropsychological Test Automated Battery). By capturing metrics related to visual memory, attention, reaction time, working memory, executive function, and semantic/verbal memory, these games offer a digital platform for evaluating senior cognitive performance. This capability not only provides valuable insights into individual progress but also allows for the standardization of data collection, potentially complementing or even augmenting traditional testing methods. The integration of such data collection capabilities into our game design allows for a



Fig. 6. Another example of a game for training attention and reaction time.

more comprehensive and objective evaluation of the games' effectiveness in maintaining and improving cognitive function in seniors. We have to note that currently we are preparing the experiments, in which we start to collect the gameplay data for the future cognitive performance assessment.

3.2 Proposed Enhancements

These enhancements fall into general improvements applicable to all games and specific improvements tailored to individual game types.

General Enhancements.

- Multi-language support, especially Slovak, with in-app language switching: This is crucial for expanding the user base and making the games accessible to a broader audience, particularly in regions where Slovak is spoken. In-app switching simplifies the process and avoids requiring users to change their device's system language.
- Option to use personalized audio recordings (e.g., MP3 files) instead of text-to-speech: Personalized recordings, such as a family member's voice, can significantly increase engagement and provide a more comforting and familiar experience for users, especially those with cognitive decline.
- User-friendly editors for creating custom game content, accessible to less tech-savvy staff: Intuitive editors will empower caregivers and staff to create tailored tasks and exercises without requiring advanced technical skills, allowing for more personalized and targeted interventions.
- Comprehensive website with application downloads, video manuals, user feedback forums, FAQs, and data encryption for sensitive content: A well-designed website is essential for distributing the applications, providing support through video tutorials and FAQs, gathering user feedback, and ensuring the secure handling of potentially sensitive user data.

- Enhanced voice output editing features (e.g., repetition, reminders): Providing more control over voice output, such as the ability to repeat instructions or set reminders, will improve the user experience and ensure clarity, particularly for users with cognitive or auditory challenges.
- Exploration of wireless button integration: Wireless buttons could offer a more accessible and flexible control mechanism for users with limited mobility or dexterity, potentially replacing or supplementing touchscreen and button interfaces.
- Multi-user versions of games for competitive play: Introducing multi-user modes will foster social interaction and motivation through friendly competition, making the games more engaging and enjoyable.
- Increased number of sample tasks included with the applications: Providing a wider range of pre-made tasks will give users a better starting point and demonstrate the capabilities of the games, encouraging exploration and use.
- Creation of sample production kits (e.g., for NFC Games: app, sample tasks, card sets): Sample kits will facilitate the adoption of the games by providing all necessary components in a convenient package, simplifying setup and implementation for users and organizations.

Game-Specific Enhancements.

- NFC Games: A protective phone case will prevent accidental button presses and screen interactions. A flat desktop enclosure will ensure consistent phone placement on the game board. An audible signal will confirm correct placement, providing feedback to the user. Integrating board game elements, such as a path to follow, will add a narrative element and increase engagement.
- Dice Games: Offering dice in various sizes will accommodate users with different hand sizes and dexterity levels. Supporting multiple dice will allow for more complex and varied game mechanics. Advanced mechanics, like colored dice or sum-based tasks, will further enhance the games' challenge and cognitive benefits.
- VidPicAskAns: Increasing the number of supported questions and answers will allow for more comprehensive assessments and learning activities. Customizable color settings will cater to individual preferences and visual needs. Multi-group play with separate answer buttons will enable collaborative and competitive scenarios.
- Attention + Response: Increasing the number of buttons will expand the range of visual field training exercises. Customizable setup options will allow the games to be tailored to individual user abilities and progress. Multi-group competitive play will add a social element and enhance motivation. Multiple test types will provide variety and target different cognitive skills.

Artificial Intelligence Integration.

Integrating AI will enhance the games' adaptability and personalization. By classifying task complexity and monitoring user performance, the AI can dynamically adjust the difficulty to maintain an optimal level of challenge, preventing frustration while maximizing learning. The AI will track user performance on individual task steps and adjust the frequency of repetition to optimize memory training, focusing on areas where the user needs the most support. By analyzing user error patterns, the AI can identify potential cognitive impairments, enabling personalized adjustments to the games and potentially prompting further assessment.

Reusability and Impact.

While primarily designed for older adults, these games have broad applicability to other age groups, people after injuries requiring rehabilitation, etc.

- **NFC Games:** These games can be used for visual field training after surgery, fine motor skill rehabilitation following injury or stroke, and cognitive training for various conditions.
- **Dice Games:** These games are effective for fine motor skill rehabilitation and cognitive training, particularly for hand-eye coordination.
- **VidPicAskAns:** This versatile tool can be used for education, reminiscence therapy, and cognitive training across different age groups and for various purposes.
- **Attention + Reaction:** These games are beneficial for training attention, reaction time, and multi-sensory perception in diverse populations.

4 Conclusions

In this paper, we presented the design, implementation, and user testing of a suite of digital cognitive games utilizing tangible physical objects, specifically tailored for seniors and individuals with limited motor skills. The invaluable feedback from potential users has refined the games' usability, accessibility, and therapeutic potential. Moreover, the games' inherent data collection capabilities, capturing metrics like reaction time and accuracy, enable quantitative performance analysis. This data offers a digital avenue for evaluating cognitive functions such as visual memory, attention, and executive function, directly comparable to standardized neuropsychological tests like CANTAB.

The successful integration of physical objects with digital gameplay has demonstrated the potential to engage seniors in a more interactive and enjoyable manner, addressing the limitations of purely mental cognitive exercises. The platform's flexibility allows for continuous expansion, incorporating new themes, objects, and challenges to ensure relevance and engagement.

Furthermore, we recognize the significant benefits of multiplayer interactions, fostering social engagement and reducing isolation. For carers, these games provide a valuable tool for monitoring cognitive function, tracking progress, and facilitating meaningful interactions, with customizable content and difficulty levels for personalized interventions.

In conclusion, our research demonstrates the potential of integrating physical objects with digital cognitive games to enhance cognitive function and quality of life for seniors. By combining user-centered design, data-driven evaluation, and flexibility, we have created a platform that provides engaging cognitive stimulation and valuable performance insights.

To further this research, we have identified key areas for future development. Longitudinal studies are needed to assess the long-term impact on user function and engagement. Integrating with telehealth platforms could facilitate remote monitoring and personalized feedback. Enhanced gamification and social features can improve motivation, while expanded AI capabilities can enable more sophisticated personalization. New game types should target specific cognitive and motor skills, and customizable hardware interfaces can accommodate diverse user needs. Integrating biofeedback mechanisms

could provide real-time engagement feedback. Ensuring cross-platform compatibility and cultural adaptation are crucial for diverse populations. Finally, collaboration with healthcare professionals will facilitate integration into rehabilitation programs and gather expert feedback. Ultimately, our aim is to contribute to the development of accessible and effective digital interventions that empower seniors to maintain their cognitive health and well-being.

Acknowledgement. The work is supported by the project No. CE0100038 DigiCare4CE of INTERREG Central Europe programme (<https://interreg-central.eu/projects/digicare4ce/>).

References

1. Nägele N, von Walter B, Scharfenberger P et al (2020) Touching services: tangible objects create an emotional connection to services even before their first use. *Bus Res* 13:741–766. <https://doi.org/10.1007/s40685-020-00114-0>
2. Skulmowski A, Pradel S, Kühnert T, Brunnett G, Rey GD (2016) Embodied learning using a tangible user interface: the effects of haptic perception and selective pointing on a spatial learning task. *Comput Educ* 92–93:64–75. <https://doi.org/10.1016/j.compedu.2015.10.011>
3. Fu Y, Hu Y, Sundstedt V (2022) A systematic literature review of virtual, augmented, and mixed reality game applications in healthcare. *Assoc Comput Mach* 3(2). <https://doi.org/10.1145/3472303>
4. Lhotska L, Husák J, Stejskal J, Kotek M, Dolezal J, Adolf, J (2022) Role of virtual reality in the life of ageing population. *Neural Netw World* 32(5):253–267. <https://doi.org/10.14311/NNW.2022.32.015>
5. Chan JYC, Liu J, Chan ATC et al (2024) Exergaming and cognitive functions in people with mild cognitive impairment and dementia: a meta-analysis. *NPJ Digit Med* 7:154. <https://doi.org/10.1038/s41746-024-01142-4>
6. Abd-alrazaq A, Abuelezz I, Al-Jafar E et al (2023) The performance of serious games for enhancing attention in cognitively impaired older adults. *NPJ Digit Med* 6:122. <https://doi.org/10.1038/s41746-023-00863-2>
7. Tuan S-H, Chang L-H, Sun S-F, Lin K-L, Tsai Y-J (2022) Using exergame-based exercise to prevent and postpone the loss of muscle mass, muscle strength, cognition, and functional performance among elders in rural long-term care facilities: a protocol for a randomized controlled trial. *Front Med* 9:1071409. <https://doi.org/10.3389/fmed.2022.1071409>
8. Lee CS, Yii W (2021) Lean rehabilitative-predictive recommender games for seniors with mild cognitive impairment: a case study. In: Gervasi O, et al (eds) *Computational Science and Its Applications – ICCSA 2021*. ICCSA 2021. LNCS, vol 12951. Springer, Cham. https://doi.org/10.1007/978-3-030-86970-0_29
9. Skikos G, Goumopoulos C (2021) A tablet-based game tool for cognition training of seniors with mild cognitive impairment. In: Jensen CS et al (eds) *Database Systems for Advanced Applications. DASFAA 2021 International Workshops*. DASFAA 2021. LNCS, vol 12680. Springer, Cham. https://doi.org/10.1007/978-3-030-73216-5_24
10. Choi E, Lee B (2025) Unlocking the potential of play: a TF-IDF analysis of ‘MapleStory’ as a serious game for cognitive enhancement in seniors. *Entertain Comput* 52:100800, ISSN 1875-9521. <https://doi.org/10.1016/j.entcom.2024.100800>
11. Guzmán DE, Rengifo CF, Guzmán JD et al (2024) Virtual reality games for cognitive rehabilitation of older adults: a review of adaptive games, domains and techniques. *Virtual Reality* 28:92. <https://doi.org/10.1007/s10055-024-00968-3>

12. Kleschnitzki JM, Beyer L, Beyer R, Großmann I (2022) The effectiveness of a serious game (MemoreBox) for cognitive functioning among seniors in care facilities: field study. *JMIR Serious Games* 10(2):e33169. <https://doi.org/10.2196/33169>
13. Skikos G, Goumopoulos C, Ferras J (2025) Combining serious games and machine learning to assess the cognitive state of seniors using the Cogniplat platform. *J Ambient Intell Human Comput* 16:303–314. <https://doi.org/10.1007/s12652-024-04946-0>
14. Abd-alrazaq A et al (2022) The effectiveness and safety of serious games for improving cognitive abilities among elderly people with cognitive impairment: systematic review and meta-analysis. *JMIR Serious Games* 10(1):e34592. <https://doi.org/10.2196/34592>
15. Zuo X, Tang Y, Chen Y, Zhou Z (2024) Review effects of electronic serious games on older adults with Alzheimer's disease and mild cognitive impairment: systematic review with meta-analysis of randomized controlled trials. *JMIR Serious Games* 12:e55785. <https://doi.org/10.2196/55785>

Young Investigators Competition



Prognostic Value of Early Interictal Epileptiform EEG Patterns in Ischemic Stroke: A Predictive Model for Functional Outcomes

Katerina Iscra¹(✉), Edoardo Ricci², Andrea Bonini¹, Giovanni Furlanis², Michele Malesani², Paola Caruso², Marcello Naccarato², Paolo Manganotti², Agostino Accardo¹, and Miloš Ajčević¹

¹ Department of Engineering and Architecture, University of Trieste, Trieste, Italy
katerina.iscra@phd.units.it

² Clinical Unit of Neurology, Department of Medical Sciences, University Hospital and Health Services of Trieste, University of Trieste, Trieste, Italy

Abstract. An accurate prediction of clinical outcomes in acute ischemic stroke is essential for improving patient management. There is growing research interest in EEG-based neurophysiological biomarkers. In this context, although the prevalence of epileptiform patterns is emerging, their role as potential predictors of outcome is still poorly investigated. In this study, we investigate whether interictal epileptiform discharges (IEDs) detected in the acute phase can contribute to predicting long-term functional outcomes in treated ischemic stroke patients and to the development of a predictive model. We retrospectively analyzed clinical, neurophysiological, and neuroimaging data from 228 acute ischemic stroke patients (mean age: 74 ± 14 years; 122 females). Patients were classified into two groups based on their 3-month modified Rankin Scale (mRS) scores: good outcomes (mRS 0–2, 127 patients) and poor outcomes (mRS 3–6, 101 patients). Standard 19-channel EEG recordings were conducted to assess epileptiform abnormalities and IEDs. Feature selection was performed based on the Gain Ratio method, and predictive models were developed using five different machine learning (ML) techniques: classification tree, logistic regression, naïve Bayes, artificial neural network, and support vector machine. Our results identified the National Institutes of Health Stroke Scale (NIHSS) at admission, mRS anamnestic, chronic heart failure, acute infectious disease, sum of lobes with a infarct area, atrial fibrillation, and IEDs as significant predictors of ischemic stroke functional outcomes. Logistic regression achieved the best performance, with an accuracy of 81% and an AUC of 0.83. These findings confirm the importance of clinical and radiological features while also highlighting the predictive value of interictal EEG epileptiform patterns for stroke outcome prediction.

Keywords: EEG · Interictal epileptiform discharges · Ischemic stroke · Machine learning

1 Introduction

Stroke is one of the leading causes of death and disability worldwide, with approximately 17 million cases reported annually according to the World Health Organization (WHO) [1]. Among the two main types of stroke, ischemic stroke accounts for approximately 85% of cases [2]. Despite advances in prevention and acute management that have improved clinical outcomes, the global burden of stroke continues to grow, made worse by population aging and the increasing prevalence of risk factors such as hypertension, diabetes, and obesity [2, 3]. Reperfusion therapies in patients with acute ischemic stroke improved functional outcomes [4]. However, some patients fail to achieve full recovery despite successful reperfusion.

One of the central challenges in stroke management is the accurate prediction of clinical outcomes to develop personalized therapeutic interventions. This prediction is particularly complex, as it depends on various clinical, radiological, and neurophysiological characteristics [5, 6]. In this context, neurophysiological biomarkers derived from electroencephalogram (EEG) are emerging as promising tools for predicting patients' functional outcome [7, 8]. Indeed, acute stroke is a neurological emergency characterized by a reduction in regional cerebral blood perfusion, leading to alterations in the brain's electrical activity [9, 10], which reflects the phenomenon of neurovascular coupling [11].

Besides EEG slowing abnormalities, increasing evidence suggests the presence of epileptiform patterns in patients with acute ischemic stroke. Recent studies have investigated the prevalence of interictal epileptiform discharges (IEDs) and seizures during the acute and subacute phases of stroke [12, 13]. In particular, status epilepticus was recorded in 71.4% of patients with periodic lateralized epileptiform discharges, reinforcing the role of EEG in identifying specific patterns and detecting purely electrographic seizures [12]. Additionally, interictal epileptiform abnormalities were observed in 25.7% of cases and showed a positive correlation with seizure recurrence [13]. However, data on interictal epileptiform discharges (IEDs) during the acute ischemic stroke phase remain insufficient, particularly regarding their predictive value in patients without clinical seizures, a history of epilepsy, or anticonvulsant treatment.

Machine learning is increasingly applied in healthcare, offering high levels of accuracy. Recent studies have developed predictive models aimed at determining functional outcomes three months after stroke [14–16]. For instance, machine learning approaches have been shown to effectively predict negative functional outcomes in patients with acute ischemic stroke [14]. Another study highlighted the importance of integrating clinical and imaging data to improve outcome predictions [15]. Furthermore, interpretable machine learning models combining clinical and radiological data available during emergency evaluations have demonstrated high accuracy in predicting three-month outcomes in patients undergoing thrombectomy for acute ischemic stroke [16].

Therefore, the aim of our study was to investigate whether the presence of the IEDs in the EEG in the first 72 h make contribute to the prediction of clinical functional outcome, as measured by the modified Rankin Scale (mRS) at three months, in treated stroke patients and produce the predictive model.

2 Methods

2.1 Study Population

We retrospectively analyzed clinical, neurophysiological, and neuroimaging data from a cohort of acute stroke patients admitted to the Stroke Unit of the Neurological Clinic at Trieste University Hospital. The inclusion criteria were the patients with ischemic stroke who underwent reperfusion treatment and had an EEG within the first three days of the sub-acute phase, irrespective of age or gender. Exclusion criteria included stroke mimics, prior epileptic seizures, ongoing anti-epileptic treatment, or logistical issues preventing EEG recording. The study was approved by the local ethics committee and conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent.

Initial evaluations included non-enhanced CT (NECT) upon admission, with some patients also undergoing CT angiography or CT perfusion scans. MRI was performed for patients without confirmed lesions on follow-up CT. Data collected included demographic details, stroke risk factors (e.g., atrial fibrillation, diabetes, hypertension, smoking, chronic heart failure), clinical features (e.g., National Institutes of Health Stroke Scale (NIHSS) on admission and discharge, mRS at discharge and 3 months), and lesion characteristics (e.g., cortical involvement, infarct location, side, sum of lobes with an infarct area, and lobe involvement).

Standardized clinical and diagnostic assessments were conducted during hospitalization and at 3 months post-stroke. Outcomes were determined via phone interviews or follow-up exams by two neurologists. Neurologic impairment was evaluated using the NIHSS, recorded on the 7th day or earlier if discharged, and functional outcomes were assessed using the mRS at 3 months. NIHSS measured stroke-related impairment across 11 items, while mRS categorized functional outcomes into seven levels (0 = no symptoms, 6 = death). Patients were classified into two groups based on their 3-month mRS scores: those with good outcomes (mRS 0–2) and those with poor outcomes (mRS 3–6).

2.2 EEG Acquisition and Analysis

Standard 19-channel (10–20 system) clinical surface EEG recordings were performed within 72 h of symptom onset using the Be Plus PRO amplifier (EB NEURO, Florence, Italy) and Ag/AgCl electrodes. EEG signals were filtered with a band-pass Butterworth filter (0.1–30 Hz). Two experienced neurologists performed a qualitative visual inspection of EEG traces following international criteria [17, 18]. Recordings included wakefulness with eyes closed, eyes open, hyperventilation, and photic stimulation maneuvers. The IEDs were identified following the International Federation of Clinical Neurophysiology guidelines as di- or tri-phasic sharp waves with a characteristic slow after wave [17].

2.3 Feature Selection, Classification and Model's Evaluation

Feature selection was performed using the Gain Ratio (GR) method, which calculates the ratio of Information Gain to the attribute's intrinsic information, reducing the bias

toward multivalued features inherent in Information Gain. This approach, particularly suitable for heterogeneous datasets like ours, assesses a feature's importance by measuring its ability to reduce uncertainty about the target variable, normalized by the intrinsic information generated during dataset splitting. This ensures that the selected features are truly informative rather than merely having a high number of distinct values. For model construction, only features with a GR value higher than 0.03 were included. Odds ratios (OR) with corresponding 95% confidence intervals (CI) were calculated for the selected features.

Five different machine learning models were considered in the study: classification tree, logistic regression, naïve Bayes, artificial neural network and support vector machine (SVM). The decision tree developed in this study is a binary tree, where each leaf contains at least two instances. It does not further split subsets with fewer than five instances and has a maximum depth of 100 levels. Additionally, node splitting stops when a 95% majority threshold is reached. The logistic regression classifier incorporates ridge regularization with a cost parameter set to 1. Instead, the Naive Bayes model includes a preprocessing step that removes empty columns, if present, and categorizes numerical values into four equal-frequency bins. The neural network consists of a single hidden layer with 100 neurons, utilizing the ReLU (Rectified Linear Unit) activation function and a stochastic gradient-based optimizer for weight adjustments. The model was trained with a maximum of 200 iterations. Finally, the SVM model was configured to minimize the error function with a cost parameter of 1. For the kernel function, which maps data into a higher-dimensional space to facilitate class separation, the Radial Basis Function (RBF) was chosen. The number of iterations was capped at 100.

The classification accuracy (CA), area under the curve (AUC), F1 score, precision, recall, and specificity were evaluated using 5-fold cross-validation. This process was repeated five times, with each subsample used once as the validation set. The overall performance metrics were calculated as the mean across all five validation folds. Additionally, confusion matrices and ROC analyses were generated for each model. All analyses were conducted using Matlab (MathWorks Inc., Natick, MA) and Python's Orange3 Data Mining library and toolbox [19].

3 Results

The study sample consisted of 228 stroke patients (74 ± 14 year, 122 females; 127 patients with good outcome and 101 patients with poor outcome) who underwent EEG within the first 72 h. The features identified as highly predictive by the GR method were NIHSS at admission (GR = 0.13), mRS anamnestic (GR = 0.05), chronic heart failure (GR = 0.05), acute infectious disease (GR = 0.05), sum of lobes (GR = 0.04), atrial fibrillation (GR = 0.03), and interictal epileptiform discharges (IEDs, GR = 0.03). This set of selected features was used to develop the machine learning models.

Table 1 reports the performance metrics of developed classification models and the ROC curves are depicted in Fig. 1. All models demonstrated medium-to-high performance, with the logistic regression model, an interpretable approach, achieving the highest accuracy. Specifically, the logistic regression model, which integrates clinical parameters and risk factors along with IEDs, predicted the outcomes of treated stroke patients with an accuracy of 81%.

Table 1. Performance measures of the five produced models.

Models	CA	AUC	F1	Precision	Recall	Specificity
Logistic Regression	0.81	0.83	0.80	0.81	0.81	0.79
SVM	0.79	0.84	0.79	0.80	0.79	0.77
Neural Network	0.79	0.84	0.79	0.79	0.79	0.77
Tree	0.77	0.77	0.77	0.78	0.77	0.75
Naïve Bayes	0.76	0.83	0.76	0.76	0.76	0.74

Table 2 presents the confusion matrix for the best-performing model, the logistic regression model. The confusion matrix demonstrates that this model correctly predicts the outcomes of 81% of stroke patients.

Table 2. Confusion matrix obtained by the best-performing model, the logistic regression.

	Good outcome	Poor outcome	Σ
Good outcome	113	14	127
Poor outcome	30	71	101
Σ	143	185	228

Finally, the association between significant features and the poor outcome identified by the logistic regression model that presented the best performance was expressed as odds ratios (ORs) with 95% confidence intervals (CIs). In particular, higher NIHSS scores at admission (OR: 1.03, 95% CI: 1.02–1.04) and greater pre-stroke functional disability, as measured by the anamnestic mRS (OR: 1.10, 95% CI: 1.04–1.15), are significant predictors of poor outcomes. Chronic heart failure (OR: 1.07, 95% CI: 0.96–1.19), atrial fibrillation (OR: 1.07, 95% CI: 0.96–1.31), and acute infectious disease (OR: 1.12, 95% CI: 0.96–1.30) showed slightly elevated odds, with confidence intervals exceeding 1. The cumulative burden of lobar pathologies (Sum of Lobs) shows a marginal association with poor outcomes (OR: 1.04, 95% CI: 1.00–1.08). Importantly, the presence of interictal epileptiform discharges (IEDs) is associated with a higher likelihood of poor outcomes (OR: 1.13, 95% CI: 1.00–1.28).

4 Discussion

One of the primary challenges in stroke care is accurately predicting clinical outcomes to enable the development of personalized therapeutic interventions. Machine learning models, based on clinical parameters, risk factors and neurophysiological biomarkers derived from EEG, hold the potential to enhance predictive accuracy and support personalized treatment planning for stroke patients.

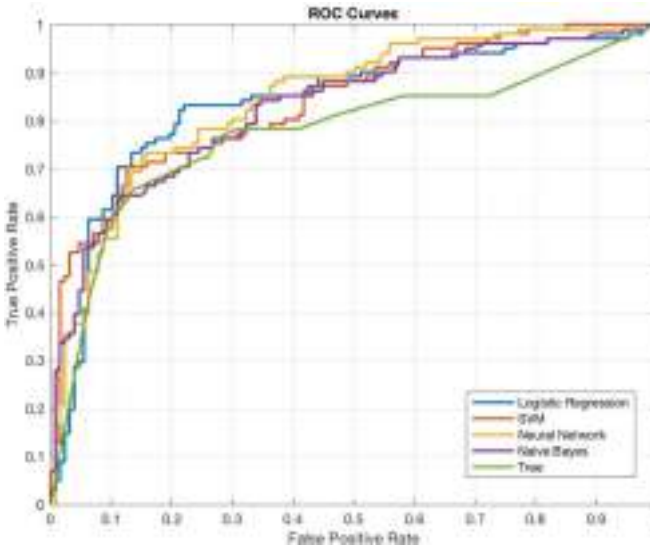


Fig. 1. ROC curves for the five produced models.

Our results highlight that the presence of IEDs, as an EEG biomarker, is a strong predictor of poor outcomes and may contribute to the prediction of functional outcome, along with other clinical and radiological features (NIHSS at admission, anamnestic mRS, chronic heart failure, acute infectious disease, sum of lobes, and atrial fibrillation) identified through feature selection.

In general, the absence of pathological EEG alterations may serve as a surrogate marker of tissue reperfusion and favorable clinical outcomes. Patients with EEG abnormalities exhibit a hypermetabolic pattern in neural tissue, resembling the hypermetabolism observed during seizures in epileptic disorders [20]. These phenomena of blood flow dysregulation, coupled with metabolic alterations, may plausibly account for a reduced functional status even in the absence of overt seizures.

Other significant parameters also showed associations with 3-month outcomes. In particular, lower NIHSS scores at admission and lower mRS anamnestic scores were strongly associated with better outcomes, confirming results obtained in different previous studies [21, 22]. Conversely, chronic heart failure and atrial fibrillation were associated with worse outcomes. The presence of these conditions is known to complicate stroke recovery and current guidelines recommend anticoagulation for chronic heart failure patients [23]. The sum of lobes affected increased the likelihood of achieving an mRS > 2. A study suggested that greater change in topology was associated with larger residual symptom burden and greater motor impairment 1, 3 and 12 months after stroke [24]. These findings highlight the predictive value of certain clinical factors, particularly NIHSS scores, pre-stroke disability, and IEDs, in identifying patients at greater risk of unfavorable outcomes.

All the models considered in this study demonstrated medium-to-high performance. The logistic regression model emerged as the best, achieving an accuracy of 81% and an

AUC value of 0.83. Notably, the top-performing model, although only slightly superior to the others, is interpretable. Such models not only provide predictive outputs but also allow for the evaluation of their plausibility by analyzing the thresholds of the considered parameters [25]. This enables a deeper understanding of which features most strongly influence the outcome and the direction of their effect, allowing users to assess and either accept or reject the model's predictions.

Several studies have employed machine learning models to predict clinical functional outcomes in stroke patients [14–16]. One study focusing on younger subjects demonstrated that mRS at admission, living alone, and a high NIHSS score at discharge were independent predictors of poor outcomes at three months, achieving good AUC values [14]. However, it should be noted that only 9% of the subjects in this study had a poor outcome, resulting in a very low positive predictive value and low F1 scores due to an imbalance between the two groups. A recent study reported that in these cases, dataset-balancing techniques should be considered [26]. Another study, involving 293 stroke patients treated with mechanical thrombectomy, developed four different machine learning models, achieving an AUC value of 0.84. The significant features identified included age, baseline NIHSS, occlusion side, degree of brain atrophy, early ischemic core, and collateral circulation deficit volume on CTA [16]. Finally, a study analyzing the three-month outcomes of 256 stroke patients used logistic regression-based machine learning models, achieving medium-to-high AUC values (ranging from 0.79 to 0.88) [15]. Our findings confirm the importance of the aforementioned clinical and radiological features while also highlighting the predictive value of interictal EEG epileptiform patterns for stroke outcome prediction.

Our study has some limitations. It is a single-center study with a moderate sample size, which may affect the generalizability of the findings to larger populations.

In conclusion, the results of this study show that the biomarkers obtained from EEG analysis, such as interictal epileptiform discharges, can contribute to the prediction of the three-month outcome of ischemic stroke patients.

Acknowledgment. The work is partially supported by the European Social Fund (ESF) and Autonomous Region of Friuli Venezia Giulia (FVG) – “Fondo Sociale Europeo Plus 2021/2027 (FSE+)”.

References

1. Doerrfuss JI, Kilic T, Ahmadi M, Holtkamp M, Weber JE (2020) Quantitative and qualitative EEG as a prediction tool for outcome and complications in acute stroke patients. *Clin EEG Neurosci* 51:121–129. <https://doi.org/10.1177/1550059419875916>
2. Caplan LR, Simon RP, Hassani S (2023) Cerebrovascular disease—stroke*. In: Zigmond MJ, Wiley CA, Chesselet M-F (eds) *Neurobiology of Brain Disorders* (Second Edition). Academic Press, pp 457–476
3. Pu L, Wang L, Zhang R, Zhao T, Jiang Y, Han L (2023) Projected global trends in ischemic stroke incidence, deaths and disability-adjusted life years from 2020 to 2030. *Stroke* 54:1330–1339. <https://doi.org/10.1161/STROKEAHA.122.040073>

4. Manning NW, Warne CD, Meyers PM (2018) Reperfusion and clinical outcomes in acute ischemic stroke: systematic review and meta-analysis of the stent-retriever-based, early window endovascular stroke trials. *Front Neurol* 9:301. <https://doi.org/10.3389/fneur.2018.00301>
5. Rajashekar D, Hill MD, Demchuk AM, Goyal M, Fiehler J, Forkert ND (2021) Prediction of clinical outcomes in acute ischaemic stroke patients: a comparative study. *Front Neurol* 12:663899. <https://doi.org/10.3389/fneur.2021.663899>
6. Duncan PW, Jorgensen HS, Wade DT (2000) Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. *Stroke* 31:1429–1438. <https://doi.org/10.1161/01.str.31.6.1429>
7. Ajčević M et al (2021) Hyper-acute EEG alterations predict functional and morphological outcomes in thrombolysis-treated ischemic stroke: a wireless EEG study. *Med Biol Eng Comput* 59:121–129. <https://doi.org/10.1007/s11517-020-02280-z>
8. Bentes C et al (2018) Early EEG predicts poststroke epilepsy. *Epilepsia Open* 3:203–212. <https://doi.org/10.1002/epi4.12103>
9. Ajčević M et al (2021) Early EEG alterations correlate with CTP hypoperfused volumes and neurological deficit: a wireless EEG study in hyper-acute ischemic stroke. *Ann Biomed Eng* 49:2150–2158. <https://doi.org/10.1007/s10439-021-02735-w>
10. Stragapede L et al (2019) Brain oscillatory activity and CT perfusion in hyper-acute ischemic stroke. *J Clin Neurosci* 69:184–189. <https://doi.org/10.1016/j.jocn.2019.07.068>
11. Rossini PM et al (2004) Does cerebrovascular disease affect the coupling between neuronal activity and local haemodynamics? *Brain* 127:99–110. <https://doi.org/10.1093/brain/awh012>
12. Mecarelli O et al (2011) EEG patterns and epileptic seizures in acute phase stroke. *Cerebrovasc Dis* 31:191–198. <https://doi.org/10.1159/000321872>
13. Abe S, et al (2022) PROPOSE study investigators. Interictal epileptiform discharges as a predictive biomarker for recurrence of poststroke epilepsy. *Brain Commun* 4:fcac312 (2022). <https://doi.org/10.1093/braincomms/fcac312>
14. Mbarek L et al (2024) Predicting 3-month poor functional outcomes of acute ischemic stroke in young patients using machine learning. *Eur J Med Res* 29:494. <https://doi.org/10.1186/s40001-024-02056-3>
15. Johnston KC, Connors AF, Wagner DP, Knaus WA, Wang X, Haley EC (2000) A predictive risk model for outcomes of ischemic stroke. *Stroke* 31:448–455. <https://doi.org/10.1161/01.str.31.2.448>
16. Jabal MS et al (2022) Interpretable machine learning modeling for ischemic stroke outcome prediction. *Front Neurol* 13:884693. <https://doi.org/10.3389/fneur.2022.884693>
17. Hirsch LJ et al (2021) American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. *J Clin Neurophysiol* 38:1–29. <https://doi.org/10.1097/WNP.0000000000000806>
18. Brigo F et al (2020) Early poststroke seizures following thrombolysis and/or thrombectomy for acute stroke: clinical and stroke characteristics. *Epilepsy Behav* 104:106353. <https://doi.org/10.1016/j.yebeh.2019.05.048>
19. Demšar J et al (2013) Orange: data mining toolbox in python. *J Mach Learn Res* 14:2349–2353
20. Struck AF, Westover MB, Hall LT, Deck GM, Cole AJ, Rosenthal ES (2016) Metabolic correlates of the ictal-interictal continuum: FDG-PET during continuous EEG. *Neurocrit Care* 24:324–331. <https://doi.org/10.1007/s12028-016-0245-y>
21. Adams HP et al (1999) Baseline NIH Stroke Scale score strongly predicts outcome after stroke. *Neurology* 53:126. <https://doi.org/10.1212/WNL.53.1.126>
22. Quinn TJ et al (2017) Pre-stroke modified rankin scale: evaluation of validity, prognostic accuracy, and association with treatment. *Front Neurol* 8:275. <https://doi.org/10.3389/fneur.2017.00275>

23. Haeusler KG, Laufs U, Endres M (2011) Chronic heart failure and ischemic stroke. *Stroke* 42:2977–2982. <https://doi.org/10.1161/STROKEAHA.111.628479>
24. Schlemm E, et al (2020) Structural brain networks and functional motor outcome after stroke—a prospective cohort study. *Brain Commun* 2:fcaa001. <https://doi.org/10.1093/braincomms/fcaa001>
25. Ahmad MA, Eckert C, Teredesai A (2018) Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. Association for Computing Machinery, New York, NY, USA, pp 559–560
26. Iskra K et al (2024) Optimizing machine learning models for classification of stroke patients with epileptiform EEG pattern: the impact of dataset balancing techniques. *Procedia Comput Sci* 246:4600–4609. <https://doi.org/10.1016/j.procs.2024.09.324>



Effect of Huntington's Disease on Electrical and Mechanical Properties of Red Blood Cells in a Mouse Model

Liliana Ponkratova¹ , E Du^{1,2} , Hongyuan Xu³, and Jianning Wei³ 

¹ Department of Biomedical Engineering, Florida Atlantic University, Boca Raton, USA
lponkratova2022@fau.edu

² Department of Ocean and Mechanical Engineering, Florida Atlantic University,
Boca Raton, USA

³ Department of Biomedical Science, Florida Atlantic University, Boca Raton, USA

Abstract. This paper presents the study of the effects of Huntington's disease (HD) on the mechanical and electrical properties of mouse red blood cells (RBCs). Electro-deformation spectroscopy (EDS), a biomechanical assay we recently developed and validated with human RBCs, is used to measure the mechanical performance of mouse RBCs, including cell deformation and fatigue resistance. Cell electrical and mechanical parameters, such as membrane permittivity, cytoplasm conductivity, and shear modulus, are further extracted using a custom MATLAB code using the frequency-dependent major and minor axes of stretched cells. Comparisons between HD and wildtype groups revealed differences in weight growth and deformation patterns with aging. HD mice struggled to gain weight, and their RBCs show lower deformability (higher membrane shear modulus) and higher cytoplasm conductivity. The findings suggest that HD alters the mechanical and electrical properties of RBCs with potential implications for tracking the disease progression.

Keywords: Huntington's disease · electro-deformation · red blood cells

1 Introduction

Huntington's disease (HD), is a genetic disorder, characterized by an expanded mutation in huntingtin gene (*HTT*) on chromosome 4. HD affects brain function and is characterized by the choreatic movement of the human body. The most accurate method for HD diagnosis is through genetic testing. The mutation is presented by an increased number of CAG repeats in *HTT*. The clinical significance of repeats can be classified into four groups. The first is with a normal range of repeats (≤ 26), the second is patients stay at risk of developing the disease (27–36), the third is pathogenic but without a full development of the disease (37–39), and the fourth is pathogenic with a full development of the disease (≥ 40) [3].

Due to the genetic mutations, not only does the central nervous system (CNS) suffer damage, but changes in the blood occur, too. Previous studies show that the disease has a

high impact on changes in the shape of red blood cells (RBCs) creating spiky-shaped cells or acanthocytes [2]. As the disease progresses, the level of the mutant huntingtin (mHTT) is measured in both peripheral blood and central spinal fluid, making it a wet marker for the progression of the disease [4]. However, assays usually require both peripheral blood and cerebrospinal fluid. Other studies show that oxidative stress markers are also useful markers, such as protein and lipid peroxidation are elevated in peripheral blood or plasma as well as in erythrocytes [5]. However, how HD affects the biomechanics of RBCs have not been determined.

The aim of our research was to characterize the mechanical performance of RBCs in a HD mouse model, utilizing a newly developed electro-deformation spectroscopy (EDS) method. Additionally, we wanted to determine whether evaluating the mechanical performance and biophysical properties of RBCs, including membrane shear modulus, cytoplasm conductivity, and membrane permittivity, can assess the impact of HD progression.

2 Biomechanical Characterization of RBCs in HD

2.1 Deformability Measurement of RBCs

RBC biophysical properties are subject to changes depending on the environment during cell development. Studies indicate that alteration in membrane proteins plays a vital role in the deformability of RBC as well as alteration in membrane lipids. SH-reagents, glycophorin-A, and monoclonal antibodies can bind to the membrane proteins and decrease cell deformability. In addition, lipid-protein interaction affects the ability of cells to deform [6]. Biomechanical characterization of RBCs can be performed using a variety of techniques [7], such as filtration, microfluidic devices, and stretching cell using external forces. Previous ultrafiltration studies showed cells from HD patients have lower deformability than those from healthy people. The results suggested that alterations to the membrane occur during its development but not due to the blood components [8]. Expression of huntingtin may take place in the bone marrow and affect the formation of red blood cells and their properties.

Another useful cell deformability measurement is through electro-deformation, which refers to the uniaxial stretching of the cells that happens under an applied electric field. Due to the electric field gradient and polarization across the membrane, cells are forced to deform. Electro-deformation, when performed using an interdigitated micro-electrode configuration in microfluidic systems, allows the measurement of multiple single cells at the same time. In the current design with $20\times$ magnification, it is possible to measure 50 to 150 cells at the same time. The method can also be used to perform fatigue testing, where cells were periodically stretched for one hour with the application of amplitude shift keying-modulated electric fields [9, 10]. Fatigue studies showed that cells were losing their deformability over time [11]. The procedure also suggests that conditions to perform fatigue can be adjusted to track the biophysical properties of RBC not only in health but also in diseases [12].

2.2 Electro-Deformation Spectroscopy

Electro-deformation spectroscopy (EDS) studies cell electro-deformation at various frequencies [10, 13]. EDS studies highlighted distinctions between human and mouse cells, such as the extent of deformation and frequency-dependence, even though mouse cells are smaller than humans [14]. Although electro-deformation and EDS are based on the same basic principle of measuring elongation of cells in electric fields, the data analysis methodologies are different. The degree of cell electro-deformation depends on its dielectric and mechanical properties. Electro-deformation allows for characterizing fundamental mechanical properties of cells using well-established methods [15, 16]. The EDS spectra allow for characterizing individual cells for both their electrical and mechanical properties by matching theoretical values with experimental measurements, as demonstrated in our recent work [10]. Three fundamental cell parameters, including membrane relative permittivity, cytoplasm electrical conductivity, and membrane shear modulus were characterized using EDS spectra of single cells.

3 Materials and Methods

3.1 Mouse Model

Initially, 13 mice were included in our studies. Three of them were wild type (WT) to maintain the experiment's control group, and ten were from a transgenic line R6/2. Six transgenic mice developed severe conditions and four mild conditions. Monthly weight records and blood samples were taken for the experiment. All animal experiments were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Institutional Animal Care and Use Committee of Florida Atlantic University (FAU). Breeding pairs of R6/2 transgenic mice (Strain #006494, [17]) were purchased from the Jackson Laboratory (Bar Harbor, ME, USA), and the line was maintained by backcrossing to CBA/J C57BL/6J F1 in the animal facilities of FAU. Animal blood was collected from facial veins of anesthetized WT and HD mouse at various time points as specified. The total duration of the experiment was 3 months since mice with the disease suffer and cannot sustain longer. During the experiment, one of the mice from the severe group died in 2 months. Five more mice were introduced to the experiment and were included in the fatigue studies after three months to better understand mouse physiology and how biophysical properties change during the animal's lifespan.

3.2 Device Fabrication and Working Medium

The microfluidic device for EDS measurements consisted of an electrode substrate and a polydimethylsiloxane (PDMS) channel. The substrate is a pre-patterned indium tin oxide (ITO) electrode array on a 0.7 mm thick glass substrate. The electrode array has a 44 μm gap and 106 μm band, allowing it to be used in the further analysis of the cells. The PDMS microchannel is 75 μm deep and has an inlet and outlet with 3 mm and 1 mm, respectively. The experimental medium was isosmotic, made of 0.3% (w/v) dextrose and 8.5% (w/v) sucrose. The conductivity of the medium was adjusted to 400 $\mu\text{S/cm}$ with phosphate-buffered saline.

3.3 EDS Experiment

To perform the EDS experiment, blood was mixed in the working medium, with a ratio of 1 μ l: 500 μ l accordingly, and loaded into the channel primed with working medium. The device was placed under the 20 $\times$ objective lens of the inverted microscope Olympus IX 73 with image contrast enhancement by a 414 ± 46 nm bandpass filter. An electric field was applied by attaching electrodes to the devices from a signal generator SIGLET SDG830. The spectroscopy was performed under constant voltage with root mean square value of 1.5 V. The included frequencies were 15 MHz, 12 MHz, 10 MHz, 8 MHz, 6 MHz to 1 MHz with a decrement of 1 MHz, and 800 kHz. Images were obtained for further analysis. The fatigue test was performed using the same blood specimen and medium. The voltage amplitude stayed the same as EDS, and the frequency was set to 5 MHz. In addition, the amplitude shift keying mode was activated, and cells were exposed to the cyclic stretching frequency of 0.25 Hz for 1 h. Images were taken before the fatigue, at intervals of 10 min, and post fatigue.

3.4 Image Analysis and EDS Characterization

The image analysis was performed in two ways with the use of ImageJ and a custom MATLAB code. Each cell was fit into an ellipse and measured major and minor diameters, a and b , respectively (Fig. 1a). Cell deformation was quantified by elliptical shift factor (ESF), the ratio of diameters of every cell in the form of $ESF = a/b$. Resistance to fatigue was measured by using ESF at 0 min (ESF_0) and ESF at 60 min (ESF_{60}) and implemented into the formula, $ESF_{60}/ESF_0 \times 100\%$. Mechanical and electrical properties, including membrane permittivity (ϵ), cytoplasm conductivity (σ), and membrane shear modulus (μ), were extracted using a custom MATLAB code based on the average EDS value per mouse in each month, as reported previously [10]. The extraction was performed for both groups for 3 months.

4 Results

Figure 1b demonstrates representative measurements of the major and minor axes of the cells during the EDS experiment. Figure 1c shows the corresponding quantification with ESF with the maximum deformation observed at 4 MHz for this blood specimen. However, it was noted in further investigation that the maximum deformation of cells occurs consistently between 3 MHz and 5 MHz for both HD and WT mice.

Figure 2 shows the changes in RBC mechanical performance with aging, evaluated by maximum cell deformation, averaging between measurements at 3 MHz and 5 MHz, as well as the resistance to fatigue. In Fig. 2a, the deformation of cells in WT decreases from month 1 to month 2, then quickly reaches a plateau with small variations along the way with their aging, while resistance to fatigue increases slightly as they mature. Resistance to fatigue represents changes in the capacity of cells to maintain their initial deformation after cycled loading. Results of the fatigue and EDS experiments suggest that RBCs in young WT mice are highly deformable, and their mechanical performance quickly become stabilized along with aging.

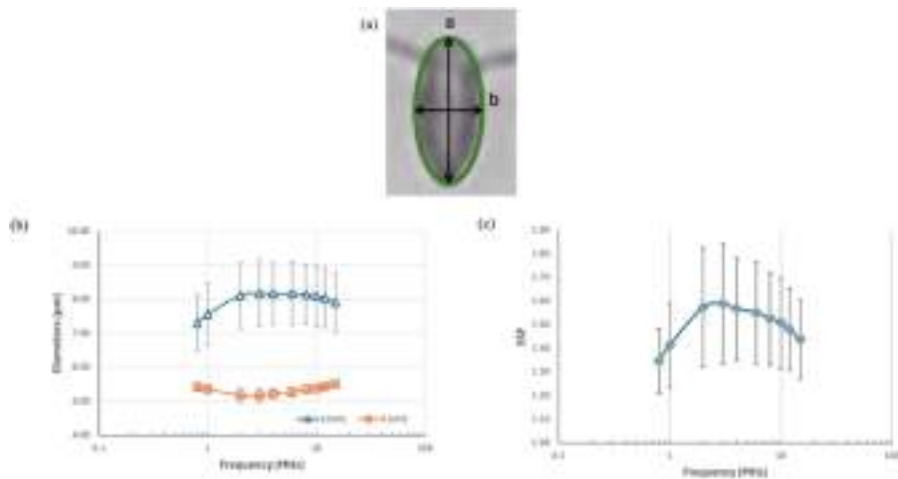


Fig. 1. Image analysis for cell deformation. (a) Fitting cell contour with an ellipse. (b) Changes in major and minor axes of cells in a WT mouse at month 2 across frequency range. (c) Corresponding ESF spectrum for quantification of cell deformation. The values presented in (b) and (c) are averages derived from 150 cells, with error bars representing the standard deviation.

In Fig. 2b, the resistance to fatigue in HD groups is higher in the first and the third months than in WT, but it is similar in the second month. In Fig. 2c, HD group follow the same pattern as WT group. Nevertheless, WT group demonstrates more deformation in the first month than HD but in the following months the deformation decreases.

Table 1 summarized the average biophysical parameters per group within the first three months, including weight, maximum deformation (ESF_{max}), mean corpuscular volume (MCV), relative permittivity of cell membrane (ϵ), cytoplasm conductivity (σ), and membrane shear modulus (μ). The value of MCV was approximated based on the major and minor axes, assuming the thickness in the third axis to be 2.5 μm . This estimation appears reasonable as the MCV values in WT are consistent with those reported for healthy mice, ranging from 45 to 55 fL [18].

Figure 3 shows the changes in three measurements that are most distinct between the two groups during the first three months, including weight, RBC cytoplasm conductivity, σ , and membrane shear modulus, μ . HD group with developing of the disease unable to gain weight, which leads to a major difference in weight change between groups. The cytoplasm conductivity is much higher in the HD group, which implies a higher intracellular ion concentrations of cells than normal RBCs.

5 Discussion

The results in Fig. 2c and Fig. 3c show related but different properties. In Fig. 2c, ESF_{max} for both groups are similar at months 2 and 3, suggesting comparable cell deformability. This result seems to conflict with a previous study by Butterfield et al. [8], which showed reduced HD RBC deformability. The discrepancy could be due to different blood sources and testing methods. Butterfield used polycarbonate filters and human RBCs to

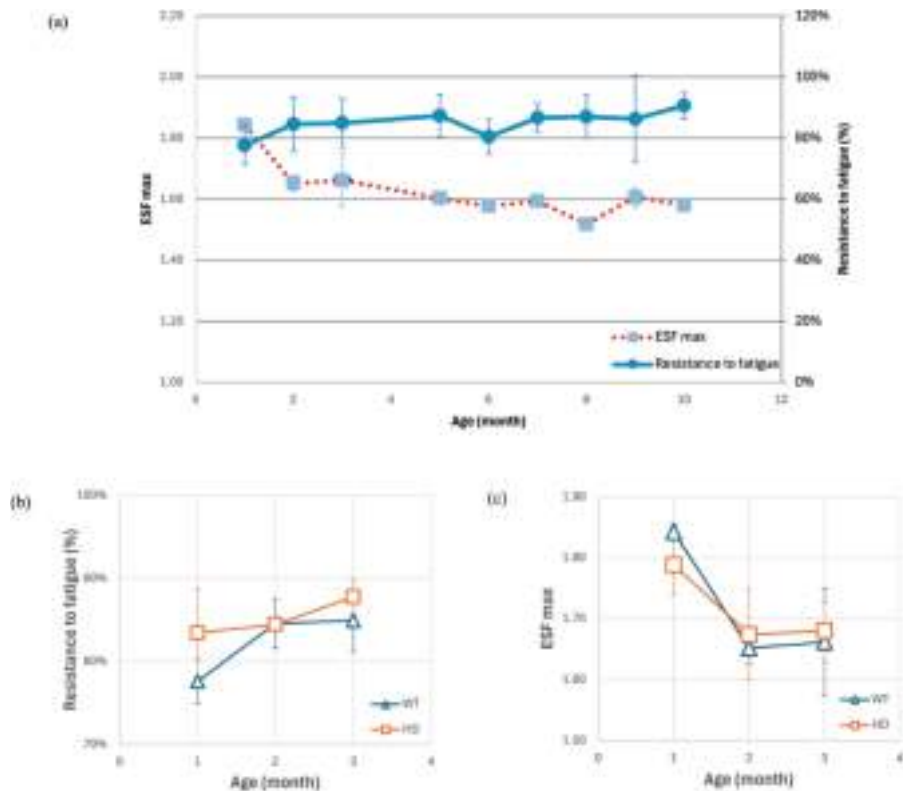


Fig. 2. Comparisons of RBC mechanical performance between HD and WT groups. (a) Maximum deformation and resistance to fatigue (a) in the WT group during mouse lifespan. (b) Resistance to fatigue between groups during the first 3 months, and (c) ESF_{max} between groups for the first 3 months. The values presented in (a), (b), and (c) are averages derived from measurements of each animal, with error bars representing the standard deviation.

Table 1. Summary table of parameters for HD and WT groups. Each parameter is represented as mean $\pm$ SD.

Age (month)	Group	Weight (g)	ESF _{max}	MCV (fL)	ϵ (μ F/m)	σ (S/m)	μ (μ N/m)
1	HD	18 $\pm$ 2.2	1.79 $\pm$ 0.06	51 $\pm$ 2.2	5.8 $\pm$ 1.3	1.29 $\pm$ 0.71	1.92 $\pm$ 0.17
	WT	18 $\pm$ 1.7	1.84 $\pm$ 0.09	52 $\pm$ 0.5	5.73 $\pm$ 0.7	0.94 $\pm$ 0.15	1.8 $\pm$ 0.05
2	HD	21.5 $\pm$ 3.1	1.67 $\pm$ 0.05	50 $\pm$ 4.5	3.59 $\pm$ 0.8	0.83 $\pm$ 0.3	1.95 $\pm$ 0.13
	WT	26 $\pm$ 2.5	1.65 $\pm$ 0.09	53 $\pm$ 3.6	3.69 $\pm$ 0.3	0.77 $\pm$ 0.07	1.91 $\pm$ 0.16
3	HD	20 $\pm$ 2.5	1.68 $\pm$ 0.08	53 $\pm$ 2.8	4.09 $\pm$ 0.8	1.1 $\pm$ 0.47	2.01 $\pm$ 0.31
	WT	28 $\pm$ 4.4	1.66 $\pm$ 0.11	52 $\pm$ 1.5	4.51 $\pm$ 1.3	0.78 $\pm$ 0.36	1.92 $\pm$ 0.55

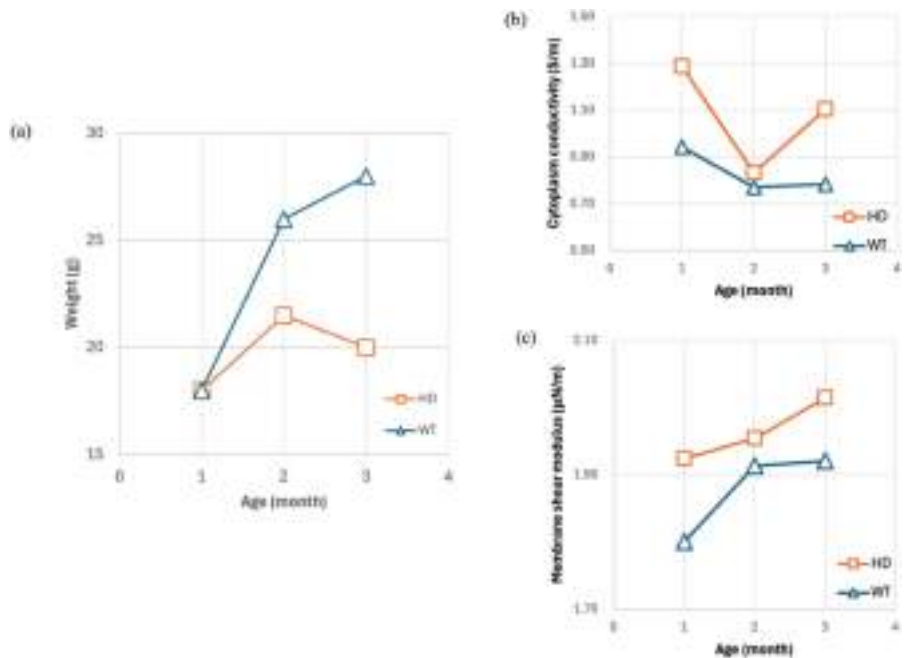


Fig. 3. The effects of HD on the measurements are evident, including mouse weight (a), RBC cytoplasm conductivity, σ (b), and RBC membrane shear modulus, μ (c). The data represent the averages for each group; error bars are excluded for scaling purposes.

measure cell deformability, which means the deformability was primarily determined by mechanical properties of cells. In contrast, the electro-deformation study was performed using mouse RBCs; the electrical force causing cell deformation depends on both dielectric and mechanical properties of cells. Assuming identical electrical properties can mask differences between the groups. EDS spectra help separate these properties, revealing the differences in the fundamental mechanical characteristics of cell membranes. Despite using average ESF and approximated MCV values, Fig. 3c shows that RBCs become more rigid with disease progression, highlighting that healthy RBCs are more deformable than those from the HD group. This is consistent with Butterfield's findings, independent of blood resources.

6 Conclusion

The study concluded that HD alters the mechanical and electrical properties of RBCs, with potential implications for tracking disease progression. Although it is not possible to determine the statistical significance of the results due to the small number of testing subjects, it is important to highlight the differences between the groups. The study presents different outcomes that can be used for future research. Firstly, measurements with single cell resolution may be useful to reveal the features that can better distinguish between HD and WT. Secondly, it indicates a correlation between aging and cell

membrane stiffness that was demonstrated by the results of resistance to fatigue and shear modulus. These measurements may provide biomarkers for monitoring disease progression. Thirdly, the differences in cytoplasm conductivity and cell deformability in the first month may be useful to detect HD in early stage before symptoms occur. Future studies should include a single cell resolution across all EDS spectrum that would allow to track each single cell and be more specific in results. It is also important to have more mice for statistical analysis and significance testing.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 1941655 and No. 2032730. Author E.D. acknowledges the support of the National Institutes of Health under Grant No. R01AI169648, and the I-HEALTH and I-SENSE pilot grants from Florida Atlantic University.

References

1. Roos RA (2010) Huntington's disease: a clinical review. *Orphanet J Rare Dis* 5:40. <https://doi.org/10.1186/1750-1172-5-40>
2. Yu, Y., Lu, Y., Wang, F., Lu, Y., Xie, B., Meng, X., & Tang, Y. Acanthocytes identified in Huntington's disease. *Frontiers in Neuroscience*, 16. (2022). <https://doi.org/10.3389/fnins.2022.913401>
3. Stoker TB, Mason SL, Greenland JC et al (2022) Huntington's disease: diagnosis and management. *Pract Neurol* 22:32–41
4. Przybyl L, Wozna-Wysocka M, Kozłowska E, Fiszer A (2021) What, when and how to measure—peripheral biomarkers in therapy of Huntington's disease. *Int J Mol Sci* 22(4):1561 (2021). <https://doi.org/10.3390/ijms22041561>
5. Túnez I, Sánchez-López F, Agüera E, Fernández-Bolaños R, Sánchez FM, Tasset-Cuevas I (2011) Important role of oxidative stress biomarkers in Huntington's disease. *J Med Chem* 54(15):5602–5606. <https://doi.org/10.1021/jm200605a>
6. Kim J, Lee H, Shin S (2015) Advances in the measurement of red blood cell deformability: a brief review. *J Cell Biotechnol* 1(1):63–79. <https://doi.org/10.3233/jcb-15007>
7. Matthews K, Lamoureux ES, Myrand-Lapierre M-E, Duffy SP, Ma H (2022) Technologies for measuring red blood cell deformability. *Lab Chip* 22(7):1254–1274. <https://doi.org/10.1039/d1lc01058a>
8. Butterfield DA, Purdy MJ, Markesbery WR (1979) Electron spin resonance, hematological, and deformability studies of erythrocytes from patients with Huntington's disease. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 551(2):452–458 (1979). [https://doi.org/10.1016/0005-2736\(89\)90020-5](https://doi.org/10.1016/0005-2736(89)90020-5)
9. Dieujuste D, Alamouti AK, Xu H, Du E (2023) Amplitude-modulated electrodeformation to evaluate mechanical fatigue of biological cells. *JoVE* (2023). <https://www.jove.com/b/65897>
10. Du E, Xu H, Ponkratova L (2024) Electro-deformation spectroscopy: a unified method for simultaneous electrical and mechanical characterization of single cells. *Acta Biomater.* <https://doi.org/10.1016/j.actbio.2024.12.012>
11. Qiang Y, Liu J, Dao M, Suresh S, Du E (2019) Mechanical fatigue of human red blood cells. *Proc Natl Acad Sci* 116(40):19828–19834. <https://doi.org/10.1073/pnas.1910336116>
12. Qiang Y, Liu J, Dao M, Du E (2021) *In vitro* assay for single-cell characterization of impaired deformability in red blood cells under recurrent episodes of hypoxia. *Lab Chip* 21(18):3458–3470. <https://doi.org/10.1039/d1lc00598g>

13. Peterlin P (2010) Frequency-dependent electrodeformation of giant phospholipid vesicles in AC electric field. *J Biol Phys* 36:339–354
14. Du E, Xu H, Wei J (2024) Electro-deformation spectroscopy of biological cells. In: 2024 IEEE 19th international conference on nano/micro engineered and molecular systems (NEMS), Kyoto, Japan (2024)
15. MacQueen LA, Buschmann MD, Wertheimer MR (2010) Mechanical properties of mammalian cells in suspension measured by electro-deformation. *J Micromech Microeng* 20(6):065007
16. Qiang Y, Liu J, Du E (2017) Dynamic fatigue measurement of human erythrocytes using dielectrophoresis. *Acta Biomater* 57:352–362
17. Mangiarini L et al (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* 87(3):493–506 (1996)
18. O’Connell KE et al (2015) Practical murine hematopathology: a comparative review and implications for research. *Comparat Med* 65(2):96–113 (2015)

Scientific Challenge Competition



Advanced Multi-class Wound Segmentation Using EfficientNet UNet

A Comprehensive Approach to Automated Wound Assessment

Athanasios Kallipolitis^(✉) , Melina Tziomaka , Argyrios Zafeiriou ,
and Ilias Maglogiannis 

Department of Digital Systems, University of Piraeus, Piraeus, Greece
{nasskall,tziomakamel,zafeiriou,imaglo}@unipi.gr

Abstract. Accurate wound image analysis is important for effective treatment planning and monitoring healing progression. This paper introduces a deep learning methodology for automated wound segmentation by developing an innovative framework that uses EfficientNet and UNet architectures. Our research tackles essential issues in wound assessment, which involve multi-class segmentation along with boundary precision and test-time augmentation. To enhance segmentation quality, the proposed methodology employs a composite loss function combining focal, boundary-aware, and dice loss modules. The experimental results show equivalent or better performance than current methodologies with our three-class model achieving an average F1 score of 0.8876. This work represents a significant advancement toward comprehensive wound assessment systems that can support decision-making across diverse healthcare settings.

Keywords: Wound Segmentation · Deep Learning · Medical Image Analysis · Computer Vision · Boundary-Aware Loss · Multi-class Segmentation

1 Introduction

Accurate wound assessment is a significant component of effective wound care due to the direct applications to treatment planning, monitoring healing progression, and improving patient outcomes. The recent advancements in computer vision and deep learning techniques have significantly enhanced the capability of analysis of medical images, thus paving the way for automated, objective, and standardized automated wound image analysis [1].

Wound segmentation is defined as the accurate definition of the wound regions from the surrounding healthy tissue or other background categories in digital images. The problem is difficult because of the heterogeneous behavior of wounds and different lighting scenarios in real life, not to mention the presence of measurement tools in wound snaps. Conventional methods rely on clinician-based assessment, which is laborious and prone to variability. Recent deep learning methods, especially convolutional neural network (CNN) inspired methods, have shown extraordinary capability in

automatically learning discriminative features and precisely segmenting the wound area while exhibiting good characteristics of objectiveness [2].

We propose a state-of-the-art wound segmentation framework that combines the expressive efficiency of EfficientNet architecture with U-Net, with the further improvement of a complex loss function. To address these issues in the wound segmentation task, the following solutions are proposed:

1. **Multi-Class Segmentation:** A single framework to make distinctions between wound tissue, background, and measurement tools (scale markers).
2. **Boundary Localization:** Utilizing a special loss function to focus on precise wound margins delineation.
3. **Class Imbalance:** Utilizing adaptive weighting techniques to mitigate the disproportionate distribution of various tissue classes.

The major contributions of this paper include the development of an EfficientNet UNet architecture for wound segmentation with variant skip connections and decoder paths, and an innovative composite loss function with a combination of focal, boundary-aware, and dice loss components to overcome class imbalance, edge accuracy, and region coherence. The proposed deep learning pipeline is an advance in automated wound image analysis, with immediate application to clinical wound documentation, measurement, and monitoring.

2 Related Work

Wound segmentation has evolved significantly over the past decade, transitioning from traditional image processing techniques to sophisticated deep learning approaches. This section provides a foundation for our work by examining key developments in the field, organized by methodological progression and application context.

2.1 Classical Approaches to Wound Segmentation

Traditional image processing has been commonly used in early wound segmentation methods. Wannous et al. [3] proposed a color and texture-based approach that uses statistical color models for classifying the wound and healthy skin. This system was accurate at a moderate level but inappropriate in different lighting conditions as well as different wound types. Similarly, Veredas et al. [4] developed a hybrid approach (region growing algorithms + neural networks) for pressure ulcer segmentation and achieved some good segmentation performance but needed significant manual intervention. Generally, these traditional approaches consisted of a pipeline of color space transformation (usually into HSV format, YCbCr, or Lab* spaces), and then thresholding or clustering methods. Although these studies performed well at detecting well-defined wounds with controlled conditions, they struggled with the heterogeneity of clinical images including those seen in practice.

2.2 Emergence of Deep Learning in Wound Analysis

Deep learning approaches for wound segmentation originated with the use of general-purpose CNN architectures. Goyal et al. [5], as one of the first methods, used a fully convolutional network (FCN) for diabetic foot ulcer segmentation, showing better results than conventional approaches. It was the work that demonstrated that deep learning can automatically extract relevant features, sparing you the time and effort of doing so with explicit feature engineering. U-Net architectures, introduced by Ronneberger et al. [6], with its focus on general biomedical image segmentation, were a significant advancement for wound segmentation. Wang et al. [7] modified U-Net to segment foot ulcers, resulting in a mean IoU of 0.79, and effectively showed how the architecture can capture both local and global information about the wounds. These early deep learning methods were usually oriented to binary segmentation (wound vs. non-wound) but did not address the need for multi-class segmentation necessary to develop comprehensive wound assessment systems incorporating measurements or different tissue types.

2.3 Specialized Architectures and Loss Functions

Being aware that general-purpose architectures cannot address all tasks, researchers started to build networks specific to the wound segmentation problem. Liu et al. [8] designed an attention-guided network that focused on boundary information using a unique boundary stream that helps to accurately delineate edges by about 5% improved performance equivalent to standard U-Net implementations. Zahia et al. followed this with a multi-scale CNN presented in [9], which operated on images of various resolutions to capture both low and high-frequency details. Residual connections were built into their architecture to promote gradient flow during training, yielding a Dice coefficient of 0.86 for the necrosis class in the pressure ulcer tissue segmentation. Several studies reviewed specialized loss functions, such as those designed to address class imbalance and boundary precision. In [10], Abraham et al. proposed a focal Tversky loss function that adaptively adjusted the weight of false positives and false negatives as the difficulty of segmentation changed. The method was especially useful in very imbalanced cases, where some classes appear many times less than others (as is often the case in medical imaging when the pathological areas are often much smaller than the surrounding healthy tissue).

2.4 Multi-class Wound Segmentation

This trend paved the way for more elaborate wound assessment systems leveraging multi-class segmentation. Oota et al. [11] proposed the WSNet architecture to identify the tissue type within wounds (granulation, slough, necrotic) achieving Dice scores ranging from 0.76 to 0.85. Their work highlighted the clinical relevance of tissue type classification for treatment planning and monitoring. At the same time, the authors from [12] worked on multi-class chronic wound classification in a novel deep learning-based model. A “Swish-ELU EfficientNet-B4” (SEEN-B4) framework tailored for multi-class wound classification, whereas most approaches concentrate on the binary classification of types of wounds. They use EfficientNet as their backbone architecture, adjusting with

Swish-ELU activation functions for improved performance. These multi-class methods were most concerned with classification of the tissue types found in the wound bed.

2.5 Transfer Learning and Data Efficiency

One of the biggest challenges, especially in medical image analysis, has been the lack of large, annotated datasets. To resolve this, various works of researchers focused on wound segmentation using transfer learning and data-efficient training. In particular, studies that utilized transfer learning methods to leverage pretrained convolutional neural networks to differentiate between diabetic foot ulcers and venous leg ulcers are especially interesting. In [13], they were able to leverage limited training data access with a test F1-score of 0.85 on cropped images, despite only training on 863 samples of wounds. The performance gap between cropped (0.85) vs. full images (0.70) shows that when it comes to transfer learning, proper region selection can really help in successfully implementing medical imaging-based applications. To combat the lack of available wound image data, researchers showed that the application of data augmentation techniques greatly improves machine learning algorithms' effectiveness in classifying wounds. In this context, reference [14] used both classical geometric transformations and more advanced techniques like Generative Adversarial Networks (GANs) to augment a small pool of wounds analysis datasets. Overall, just these geometric approaches alone were found to increase F1 scores by as much as 11%, and using GANs their realistic wound generation was achieved with larger variation. Even with these improvements in data efficiency, most methods still relied on uniform annotation strategies over the complete training dataset. A main goal of this approach was to learn from heterogeneously annotated datasets, such as those in which some images have full multi-class annotation and others that have only part of them annotated, a challenge that remained largely unaddressed.

2.6 Relationship to Our Work

The work presented herein elaborates on these past advances while also filling a number of persistent voids in the literature. To the best of our knowledge, our work is the first to propose a unified framework consisting of architecture designs and loss functions that result in a strong multi-class wound segmentation system, rather than being limited to architectural innovations or application-specific optimizations.

We use the EfficientNet UNet architecture which grasps the effectiveness of U-Net designs while using the efficiency and extraction of features done by EfficientNet [15], similar to the previous transfer learning techniques explored. This boundary-aware loss component builds on the work of Liu et al. [8] on boundary enhancement.

Together, these capabilities, compiled into an integrated system that can segment wounds and measurement tools simultaneously, represents an evolution of the current state of the art toward tools for comprehensive wound assessment that can aid clinical decision-making along the wound care continuum.

3 Methodology

This section presents a detailed overview of the methodological approach developed for automated wound segmentation.

3.1 Baseline Architecture: EfficientNet UNet

Our baseline approach utilizes an EfficientNet UNet architecture, which combines the efficiency and powerful feature extraction capabilities of EfficientNet with the robust segmentation performance of the U-Net design (Fig. 1). For concluding the use of the EfficientNet architecture as a feature extractor, we conducted extended comparative experimentations using several competent and widely used feature extractors. The EfficientNet achieved higher scores compared to ResNet34 [20], ResNet50, Swin Transformer [21], and DenseNet121 [22] encoder-decoder architectures. We report the results of these experiments.

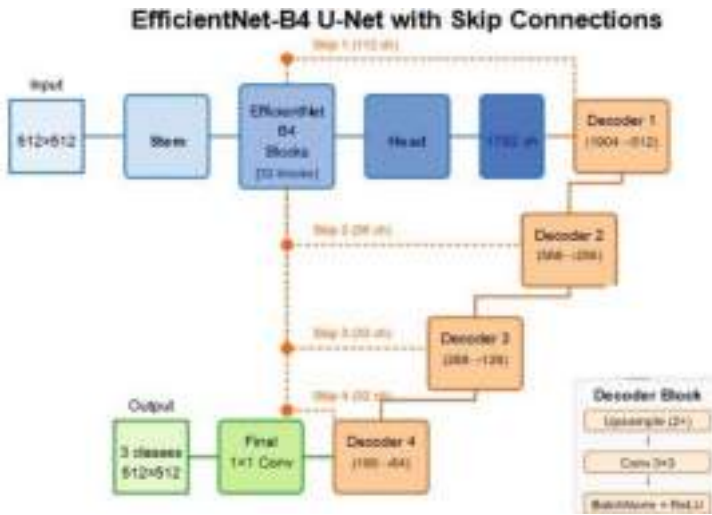


Fig. 1. UNet architecture with skip connections and an EfficientNetB4 backbone pretrained on ImageNet.

Encoder Design. The encoder component leverages a pretrained EfficientNet-B4 backbone, which provides several advantages for medical image analysis. The pre-training on ImageNet enables the network to transfer general visual recognition capabilities to the specific domain of wound assessment. This transfer learning approach is particularly valuable in medical imaging where annotated data may be limited. The network extracts features at strategic points (indices 3, 5, 9, 15) of the EfficientNet blocks, creating multi-scale representations that capture both fine details and broader contextual information. These skip connections maintain features at different levels of abstraction with channel dimensions of 112, 56, 32, and 32 respectively, preserving spatial information that would otherwise be lost through downsampling operations.

Decoder Architecture. The decoder contains four upsampling blocks that are stacked in an efficient manner to successively gain spatial resolution at the cost of reducing feature channel dimensionality. Each block and its operations follow a consistent pattern of operations: bilinear upsampling by a factor of 2, followed by a 3×3 convolution with

padding, batch normalization, and ReLU activation. His process allows the network to incrementally recover the spatial characteristics of the input image, while also combining the contextual information from the encoder pathway. The first decoder block takes as input 1904 channels and generates 512 channels, the bridge between the deepest and most spatially resolved representations. Input 568 channels are compressed down to 256 output channels in the next block. In the third block, the output dimension is further reduced to 128 through a mapping of 288 input channels, while the fourth block finalizes the spatial down-sampling by converting 160 channels to 64 output channels. This gradually diminishing size of feature dimension helps the network narrow down the critical information for the segmentation task. The output layer is a 1×1 convolution, which maps our feature maps to three output classes (wound, background, scale marker) at the target resolution of 512×512 pixels. This architecture balances the trade-off between detailed feature representation and computational cost, an important consideration for potential clinical deployment.

Loss Function Design. A critical innovation in our approach is the development of a composite loss function that addresses multiple segmentation challenges simultaneously. The combined loss function (Eq. 1) is formulated as (Fig. 2):

$$\text{Combined Loss} = \alpha * \text{Focal Loss} + \beta * \text{Boundary Loss} + (1 - \alpha - \beta) * \text{Dice Loss} \quad (1)$$

where $\alpha = 0.3$ and $\beta = 0.3$, balancing the contribution of each component. This weighted combination enables the model to simultaneously optimize for different aspects of segmentation quality.

The Focal Loss component (Eq. 2) addresses class imbalance by focusing on hard-to-classify examples, using a focusing parameter $\gamma = 2.0$. This formulation reduces the relative contribution of well-classified examples and enhances the importance of challenging cases, which is particularly valuable for wound boundaries and small features that might otherwise be overlooked. By dynamically adjusting the loss weight based on prediction confidence, the focal loss helps prevent the network from being dominated by easily classified background pixels.

$$\text{FocalLoss} = -(1 - p_t)^\gamma * \log(p_t) \quad (2)$$

The Boundary-Aware Loss component (Eq. 3) emphasizes precise delineation of wound edges by applying an exponential weighting based on the gradient magnitude of ground truth masks. This component is formulated as:

$$\text{Boundary Loss} = \text{NLL}(x, y) * e^{\theta * |\nabla y|} \quad (3)$$

where NLL represents the Negative Log Likelihood and $\theta = 1.0$ controls the boundary emphasis strength. The gradient magnitude $|\nabla y|$ is higher at the boundaries between different classes, causing the loss to be weighted more heavily in these regions. This encourages the network to pay special attention to boundary precision, which is clinically important for accurate wound measurement and assessment.

The Dice Loss component (Eq. 4) optimizes for overlap between predictions and ground truth, improving segmentation coherence:

$$Dice\ Loss = 1 - \frac{2 * |X \cap Y| + \varepsilon}{|X| + |Y| + \varepsilon} \quad (4)$$

A smoothing factor $\varepsilon = 1e-6$ ensures numerical stability. The Dice coefficient directly relates to the quality of spatial overlap, making it an intuitive metric for segmentation tasks. By incorporating it into the loss function, we encourage the network to maximize region-level accuracy rather than just pixel-level classification.

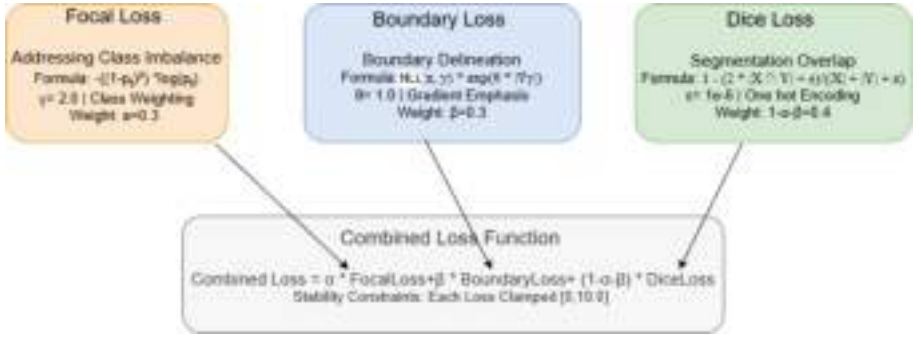


Fig. 2. Calculation of overall loss function

Class Weighting Strategy. To address the inherent class imbalance in wound images, we implement an adaptive class weighting approach. The strategy applies inverse frequency weighting, where weights are calculated as 1.0 divided by the class counts (with a small epsilon added for numerical stability). The wound class weight is multiplied by a factor of 1.5, reflecting the clinical importance of accurate wound delineation. All weights are normalized to ensure they sum to 1.0, maintaining a balanced learning signal.

Test-Time Augmentation for Robust Wound Segmentation. Our segmentation pipeline employs Test-Time Augmentation (TTA) as a post-processing strategy [16] to enhance prediction robustness. TTA applies multiple transformations to each test image, generates predictions for each variant, and aggregates these predictions into a final segmentation mask. The TTA module utilizes eight transformations: original image, horizontal flip, vertical flip, 90° and 270° rotations, brightness/contrast adjustment ($\pm 10\%$), and slight zoom variations (scale factors 1.05 and 0.95). For spatially altering transformations, predictions are realigned to the original image space before aggregation. Predictions are combined using channel-wise averaging, as shown in Eq. 5:

$$P_{final}(x, y, c) = \frac{1}{N} \sum_{i=1}^N P_i(x, y, c) \quad (5)$$

$N = 8$ is the number of transformations, and $P_i(x, y, c)$ is the SoftMax probability for class c at position (x, y) from the i -th augmentation. The final segmentation mask is determined by selecting the class with highest probability at each pixel (Eq. 6):

$$M_{final}(x, y) = \operatorname{argmax}_c P_{final}(x, y, c) \quad (6)$$

Our validation shows TTA consistently improves segmentation performance, with mean Dice score improvements across all classes compared to standard inference. The wound class benefits from a slight improvement, providing clinically relevant enhancements to boundary precision and handling of challenging cases with variable imaging conditions.

3.2 Ensemble Modeling

To further enhance the robustness and accuracy of our wound segmentation framework, we explored ensemble modeling by combining predictions from two distinct architectures: the EfficientNet UNet model (described in Sect. 3.1) and a BEiT-based segmentation model. Ensemble methods leverage the complementary strengths of individual models, potentially improving generalization and performance on challenging cases, such as those with complex wound boundaries or variable imaging conditions.

Ensemble Architecture. The ensemble model integrates two backbone architectures: the EfficientNet UNet and a BEiT-based segmentation model. The EfficientNet UNet, detailed in Sect. 3.1, combines an EfficientNet-B4 encoder pretrained on ImageNet with a U-Net decoder featuring skip connections, excelling at multi-scale feature extraction. In contrast, the BEiT (BERT Pretraining of Image Transformers) [18] architecture, inspired by the BERT (Bidirectional Encoder Representations from Transformers) [19] model from natural language processing, leverages a masked token prediction strategy to learn robust image representations. We adapted BEiT for semantic segmentation, configuring it with a patch size of 16 and an input resolution of 512×512 . To achieve the target 512×512 resolution, we modified BEiT with a custom decoder comprising four transposed convolutional layers to upsample feature maps from a 32×32 spatial grid. Additionally, a Conditional Random Field (CRF) layer was appended to enhance spatial coherence in the predictions of both models.

The ensemble combines predictions from these models by averaging their softmax outputs across the three output classes (wound, background, and scale marker). Formally, for an input image I , the ensemble prediction $P_{ensemble}$ is computed as:

$$P_{ensemble}(x, y, c) = \frac{1}{N} \sum_{i=1}^N P_i(x, y, c) \quad (7)$$

where $P_i(x, y, c)$ is the SoftMax probability for class c at pixel (x, y) from the i -th model, and N is the number of models in the ensemble (here, $N = 2$ for the full ensemble). The final segmentation mask is derived by selecting the class with the highest averaged probability at each pixel:

$$M_{ensemble}(x, y) = \operatorname{argmax}_c [P_{ensemble}(x, y, c)] \quad (8)$$

4 Experimental Setup and Results

This section presents the experimental configuration and performance evaluation of our wound segmentation approaches. We describe the datasets utilized, the implementation details, and comprehensive results for our baseline EfficientNet UNet model.

4.1 Dataset

Competition Dataset: A specialized dataset of 371 images was used for the development and evaluation of measurement tool (scale marker) segmentation capabilities. This dataset [17] uniquely provides three-class annotations distinguishing between wound regions, background tissue, and measurement tools.

Dataset Composition and Preparation: For the baseline model, the training competition dataset was utilized and divided as follows:

Training Set: 263 images (75% of training competition data).

Validation Set: 87 images (25% of training competition data).

Test Set: 80 images.

All images were standardized to 512×512 -pixel resolution while preserving the aspect ratio through appropriate padding. Color normalization was applied consistently across all datasets to account for variations in imaging equipment and lighting conditions.

4.2 Implementation Details

Hardware and Software Environment. All experiments were conducted using the following configuration:

- **Hardware:** 2 NVIDIA GeForce RTX 4090
- **Framework:** PyTorch 2.6.0
- **Training Environment:** Ubuntu 24.04.2 LTS, CUDA 12.8, cuDNN 9.1.0

Baseline Model Configuration and Training. The training process for all models of Sect. 3.1 followed these specifications:

- **Batch Size:** 16 images per GPU
- **Epochs:** Up to 50, with early stopping based on validation performance
- **Gradient Accumulation:** 2 steps (effective batch size of 32)
- **Optimizer:** AdamW with parameters as described in the Methodology section
- **Memory Management:** Mixed precision training (FP16/FP32) to optimize GPU memory usage

The baseline EfficientNet UNet model was trained with a well-defined procedure to ensure adequate convergence and performance. We used the EfficientNet-B4 as the encoder, initializing it with weights pretrained on ImageNet, which contains more than 14 million natural images divided into 1,000 categories, enabling rich feature extraction capabilities. This knowledge transfer has enabled the network to acquire a strong starting

point for learning general visual features that can be adjusted to the relevant properties of wound images.

Specifically, we first trained the model with a frozen encoder (weights frozen) for 10 epochs with a learning rate of $5e-4$. For the first phase, the decoder layers were frozen, and this allowed the randomly initialized decoder to adjust to the encoder's feature representations trained in the previous step. This prevents slippage of the useful knowledge in the encoder. This strategy sped convergence by restricting initial learning to the task-specific segmentation components.

After this first phase, all network parameters (encoder and decoder) were unfrozen and fully fine-tuned in the subsequent epochs. They were built so that the encoder could retain pre-training benefits while enabling the entire network to adapt to the intricacies of the segmentation tasks associated with wounds. The encoder gradually specialized its feature extraction to wound characteristics.

A learning-rate reduction strategy was also employed during the training, which monitored validation loss and decreased the learning rate when performance plateaued. This enabled the model to take broader optimization steps in early training but make more minute changes once closer to convergence to prevent oscillations around local minima. We implemented model checkpointing based on validation Dice score during the training process to save the weights of the model displaying the best generalization performance rather than using the final model after a set number of epochs. The best-checkpoint approach mitigates potential overfitting and captures the model at optimal generalization.

Ensemble Modeling Configuration and Training. Both models in the ensemble were trained independently on the full competition training dataset (371 images) using the same preprocessing and augmentation pipeline as the baseline model (Sect. 3.1), employing the Adam optimizer and a cross-entropy loss with inverse frequency class weighting to address class imbalance. The EfficientNet UNet model was trained with a learning rate of $5e-4$, while the BEiT-based model used a learning rate of $5e-5$, both running for up to 500 epochs with early stopping based on validation performance. The best model weights were saved based on the combined F1 score, defined as (Eq. 9):

$$F1_{combined} = 0.75 \times F1_{wound} + 0.25 \times F1_{meter} \quad (9)$$

4.3 Experimental Results

Baseline Model Performance. The baseline EfficientNet UNet model achieved strong performance on the wound segmentation task, as evidenced by the following quantitative results on the validation set created by 25% of the training competition dataset (Table 1). When examining the baseline model's performance metrics on the training competition dataset, the Background class is segmented with a Dice score of 0.9948, while the Scale Marker class is segmented with high accuracy as indicated by the 0.9723 Dice score. For the typically more challenging Wound class, a strong Dice score of 0.9374 is attained. By carefully reviewing the curves of the training process of the baseline models as shown in Figs. 3 and 4, a highly successful outcome for the skin lesion segmentation model. The loss curves demonstrate smooth convergence without overfitting, with both training

and validation loss steadily decreasing and stabilizing around epoch 20. Performance metrics reflect satisfactory results, with the final mean IoU being near 0.94 and the Dice score above 0.96.

Table 1. Class-Specific Performance Metrics for Baseline Model on 25% of the training competition dataset.

Class	IoU	Dice
Wound	0.8825	0.9374
Background	0.9897	0.9948
Scale Marker	0.9461	0.9723
Mean	0.9394	0.9682

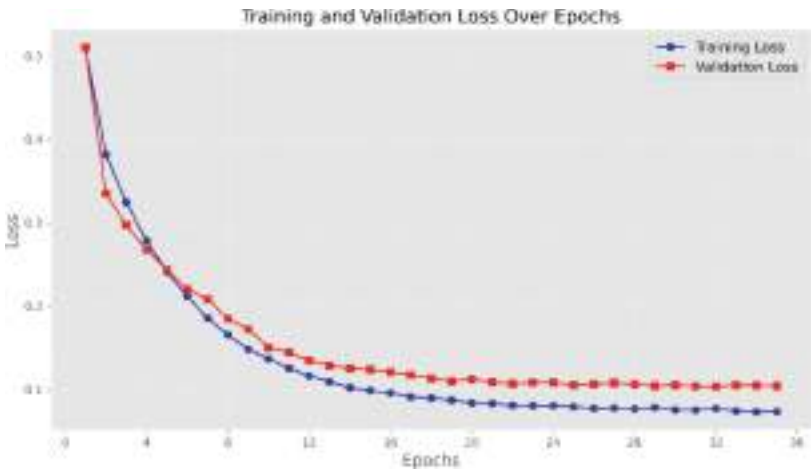


Fig. 3. Loss curves for the baseline model on the training competition dataset.

The baseline model was evaluated on the validation competition dataset (Table 2) for the F1 scores. The positive findings are further supported by the competition evaluation metrics, where a Scale Marker F1 score of 0.9645 is reached, and a strong Wound F1 score of 0.8705 is documented. A combined F1 score of 0.8940 and mean F1 score of 0.8855 are calculated, confirming consistent high performance across the entire validation image set.

Ablation Studies. To understand the contribution of different components in our methodology, we conducted ablation studies with the following variants:

- 1. **Feature extraction architectures:** Various feature extractors were tested in place of the EfficientNet and performed less competently on the validation competition dataset. The specific scores are reported on Table 3
- 2. **Loss Function Components:** Models trained with different combinations of loss terms showed that:

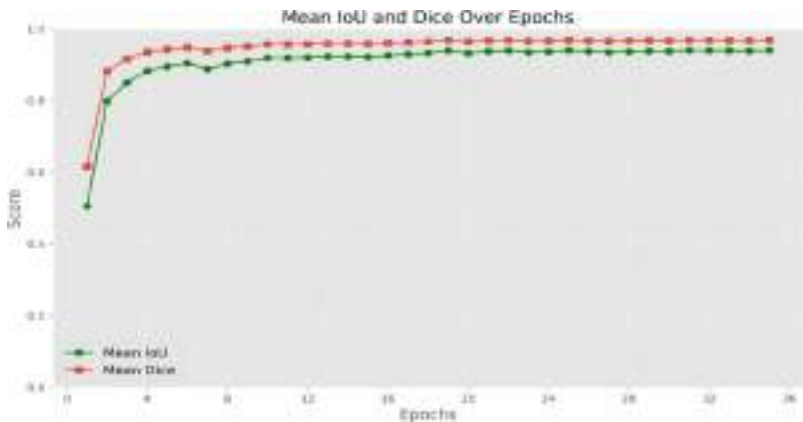


Fig. 4. Mean IoU and Dice curves for the baseline model on the training competition dataset.

Table 2. Competition Evaluation Metrics for Baseline Model on the validation competition dataset

Metric	Value
Wound F1 Score	0.8705
Scale Marker F1 Score	0.9645
Combined F1 Score	0.8940
Mean F1 Score across images	0.8855

- o Removing the Boundary-Aware Loss decreased the wound Dice score by 3.2 percentage points
- o Removing the Focal Loss resulted in a 2.1 percentage point decrease
- o Removing the Dice Loss component led to a 1.8 percentage point decrease

To further improve the segmentation results, the TTA module was added to the machine learning as described in the methodology section. This add-on slightly enhances the final results as shown in Tables 3 and 4.

In Fig. 5, examples of generated segmentation masks for the developed model are shown. The right column shows the original images, the middle column shows the ground truth, and the left column shows the generated masks. Overall, the model demonstrates good capability in segmenting both wounds and measurement tools across different wound sizes and contexts, though with varying levels of precision in boundary delineation. Performance appears strongest on larger wounds with clear boundaries and more challenging on smaller wounds or those in complex anatomical regions. When compared to the generated masks with TTA as showcased in Fig. 6, little difference can be verified with the human eye (Table 5).

As indicated by the competition committee, an executable application was developed to accept input from a fixed path and output generated masks in a predefined output folder for further evaluation of the segmentation model. The application provides feedback of

Table 3. Ablation study on the competition validation dataset for the feature extractor.

Architecture	Class	IoU	Dice
ResNet34	Wound	0.5482	0.7082
	Background	0.9358	0.9669
	Scale Marker	0.8784	0.9353
ResNet50	Wound	0.6998	0.8234
	Background	0.9625	0.9809
	Scale Marker	0.8891	0.9413
Swin Transformer	Wound	0.7276	0.8423
	Background	0.9421	0.9702
	Scale Marker	0.6902	0.8167
DenseNet121	Wound	0.7917	0.8837
	Background	0.9735	0.9866
	Scale Marker	0.8933	0.9436

Table 4. Class-Specific Performance Metrics for Model with TTA on 25% of the training competition dataset

Class	IoU	Dice
Wound	0.8894	0.9413
Background	0.9904	0.9952
Scale Marker	0.9486	0.9736
Mean	0.9428	0.9700

Table 5. Competition Evaluation Metrics for Model with TTA on the validation competition dataset

Metric	Value
Wound F1 Score	0.8753
Scale Marker F1 Score	0.9685
Combined F1 Score	0.8986
Mean F1 Score across images	0.8876

the ongoing operations concerning the image processing and mask prediction and saves the generated masks as a.png file with the same name as the input images. The predicted mask is composed only of 3 colors: black, i.e. $RGB = (0, 0, 0)$, for wounds, gray, i.e. $RGB = (127, 127, 127)$ for background, and white, i.e. $RGB = (255, 255, 255)$ for

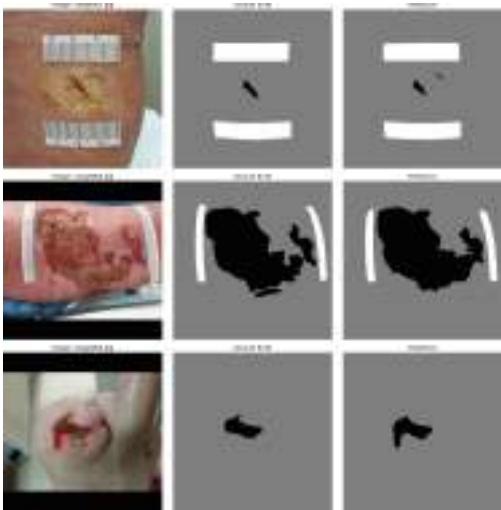


Fig. 5. Results from the application of EfficientNet UNet on the competition validation data

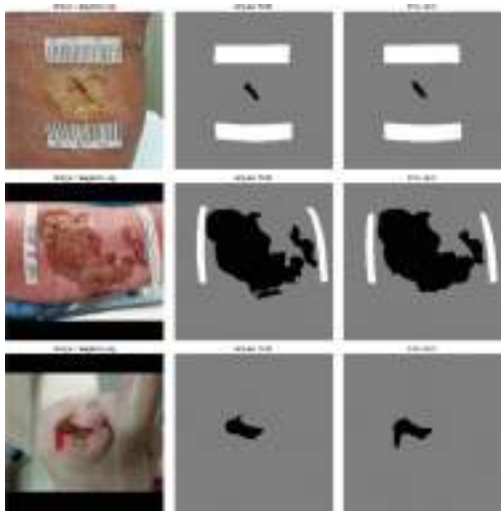


Fig. 6. Results from the application of EfficientNet UNet with TTA on the competition validation data

scale markers. For the provided test with 98 images, the overall F1 score was 0.8584 for wounds, and 0.9665 for scale markers.

Ensemble Modeling. The experimental evaluation of our ensemble approach revealed significant improvements in segmentation performance across all three classes—wound, background, and scale marker—compared to the individual models. The combination

of EfficientNet UNet and BEiT in the full ensemble consistently outperformed the standalone architectures, achieving a combined F1 score that surpassed both the EfficientNet UNet's 0.9193 and BEiT's 0.9256, reaching 0.9292 on the validation set. This enhancement reflects the ensemble's ability to integrate the convolutional precision of EfficientNet UNet with the global contextual understanding of BEiT, resulting in more accurate wound boundary delineation and robust scale marker detection. Notably, the wound class, which poses the greatest challenge due to its heterogeneity, saw a marked increase in F1 score, underscoring the clinical relevance of the ensemble strategy. Tables 6 and 7 present the IoU/Dice and F1 scores, respectively, for EfficientNet-B4, BEiT, and their average ensemble. In Fig. 7, examples of segmentation masks generated by the ensemble model (EfficientNet UNet+BEiT) are presented. The right column displays the original images from the competition test set, the middle column shows the corresponding ground truth annotations, and the left column illustrates the masks predicted by the ensemble.

Table 6. Class-Specific Performance Metrics for the Ensemble Model on the Competition Test Set

Model/Class	IoU	Dice
EffNetUNet/Wound	0.8277	0.9058
BEiT/Wound	0.8428	0.9147
Ensemble/Wound	0.8501	0.9190
EffNetUNet/Background	0.9799	0.9898
BEiT/Background	0.9811	0.9905
Ensemble/Background	0.9820	0.9909
EffNetUNet/Scale Marker	0.9232	0.9601
BEiT/Scale Marker	0.9199	0.9583
Ensemble/Scale Marker	0.9280	0.9627

The ensemble model while demonstrating higher performance in the training samples, failed to excel in the additional test set as provided by the competition committee and returned an overall F1 score of 0.8847.

5 Discussion

This study advances automated segmentation of wounds using an EfficientNet backbone implemented within a U-Net-like architecture and accompanied by a custom loss function and TTA. It also successfully tackles the challenges of multi-class wound segmentation, achieving high segmentation accuracy between wound tissue, background, and measurement tools.

The strong performance metrics achieved by our models (mean Dice score of 0.9700 on training data and mean F1 score 0.8876 on validation data) demonstrate the efficacy of the model. In fact, the heterogeneous nature of wound tissue makes the wound

Table 7. Competition Evaluation Metrics for the Ensemble Model on the Competition Test Set

Model/Metric	Value
EffNetUNet/Wound F1 Score	0.9057
BEiT/Wound F1 Score	0.9147
Ensemble/Wound F1 Score	0.9188
EffNetUNet/Scale Marker F1 Score	0.9601
BEiT/Scale Marker F1 Score	0.9583
Ensemble/Scale Marker F1 Score	0.9605
EffNetUNet/Combined F1 Score	0.9193
BEiT/Combined F1 Score	0.9256
Ensemble/Combined F1 Score	0.9292

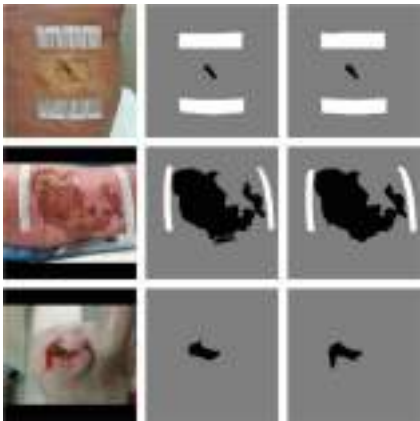


Fig. 7. Segmentation Masks Generated by the Ensemble Model (EfficientNet UNet+BEiT).

class segmentation performance (F1 score 0.8753) particularly significant. Combined with clinical history, examination, and treatment, these findings will have direct clinical implications by reducing inter-observer variability in assessments.

Our ablation studies determined that the boundary-aware loss component contributed most significantly to performance (removing this loss component resulted in a 3.2% decrease in wound Dice score). This highlights the importance of accurate boundary delimitation in wound analysis (i.e., accurate borders influence clinical measurements). The simultaneous solution to multiple aspects of the segmentation problem with our composite loss function—consisting of the focal, boundary-aware, and Dice components—was more successful than focusing on any one of the loss functions alone. The class weighting technique, with a notable focus on the wound class, effectively addressed the learning imbalance irrespective of the class imbalance within the images.

On the other hand, there are still limitations to this promising result. Validation over a more heterogeneous wound population could increase the generalizability of our approach. While it decreased the computations required to train from scratch by leveraging EfficientNet, this still raises engineering challenges for implementation in resource-limited settings. Difficult cases include complex wound margins with undermining and/or tunneling, and subtle color transitions. Furthermore, our approach here only treats segmentation as a standalone task although clinical decision-making involves temporal data integration to understand healing trajectory.

6 Conclusion

In this paper, we have proposed a deep learning model for segmentation of the wound, emphasizing the range integration of the EfficientNet UNet models. We also demonstrated the effectiveness of the EfficientNet UNet architecture for multi-class wound segmentation as evidenced by high accuracy for differentiating wounds from healthy and background tissue and measurement tools while preserving computational efficiency.

In order to overcome the unique challenges in wound segmentation with a specific focus on boundary delineation, we created a composite loss function with a combination of focal, boundary-aware, and Dice loss components and applied TTA as a post processing step to refine segmentation prediction.

The paper proposes a system that could improve clinical practice through automating wound measurements, ensuring consistency of documentation, and providing reliable monitoring of healing progress. The combination of scale marker detection and wound segmentation simplifies the documentation procedure, and the computational speed of these models allows them to be implemented for point-of-care devices.

When mature and validated further, these technologies have the potential to change the practice of wound care. Second, the incorporation of novel wound segmentation into clinical workflows may bring standardization of documentation, enhance allocation of resources, and deliver a more personalized treatment. Continued collaboration among computer scientists, biomedical engineers, and wound care specialists will be vital to translating these research innovations into meaningful clinical impact.

Acknowledgements. This work has been partly supported by the University of Piraeus Research Center.

References

1. Zhang R, Tian D, Xu D, Qian W, Yao Y (2022) A survey of wound image analysis using deep learning: classification, detection, and segmentation. *IEEE Access* 10:79502–79515
2. Anisuzzaman DM, Wang C, Rostami B, Gopalakrishnan S, Niezgoda JA, Yu Z (2020) Image based artificial intelligence in wound assessment: a systematic review. In: *Advances in wound care*
3. Wannous H, Treuillet S, Lucas Y (2007) Supervised tissue classification from color images for a complete wound assessment tool. In: *2007 29th Annual international conference of the IEEE engineering in medicine and biology society*, pp 6031–6034

4. Veredas FJ, Mesa H, Morente L (2010) Binary tissue classification on wound images with neural networks and Bayesian classifiers. *IEEE Trans Med Imaging* 29:410–427
5. Goyal M, Reeves ND, Rajbhandari SM, Ahmad N, Wang C, Yap M (2019) Recognition of ischaemia and infection in diabetic foot ulcers: dataset and techniques. *Comput Biol Med* 117:103616
6. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. *arXiv:abs/1505.04597*
7. Wang C et al (2015) A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. In: 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2415–2418 (2015)
8. Liu Y, Cheng M, Hu X, Wang K, Bai X (2016) Richer convolutional features for edge detection. *IEEE Trans Pattern Anal Mach Intell* 41:1939–1946
9. Zahia S, Sierra-Sosa D, Garcia-Zapirain B, Elmaghaby AS (2018) Tissue classification and segmentation of pressure injuries using convolutional neural networks. *Comput Methods Programs Biomed* 159:51–58
10. Abraham N, Khan NM (2018) A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp 683–687
11. Oota SR, Rowtula V, Mohammed S, Liu M, Gupta M (2023) WSNet: towards an effective method for wound image segmentation. In: 2023 IEEE/CVF winter conference on applications of computer vision (WACV), pp 3233–3242 (2023)
12. Aldoulah ZA, Malik H, Molyet R (2023) A novel fused multi-class deep learning approach for chronic wounds classification. *Appl Sci* 13:11630. <https://doi.org/10.3390/app132111630>
13. Malihi L et al (2022) Automatic wound type classification with convolutional neural networks. *Stud Health Technol Inform* 295:281–284
14. Narayanan H, Ghanta S (2024) A study of data augmentation techniques to overcome data scarcity in wound classification using deep learning. *arXiv:abs/2411.02456*
15. Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. *arXiv:abs/1905.11946*
16. Gaillochet M, Desrosiers C, Lombaert H (2023) TAAL: test-time augmentation for active learning in medical image segmentation. *arXiv:abs/2301.06624*
17. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43:17–29. <https://doi.org/10.1016/j.bbe.2022.11.004>
18. Bao H, Dong L, Wei F (2021) BEiT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*. <https://arxiv.org/abs/2106.08254>
19. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://arxiv.org/abs/1810.04805>
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
21. Liu Z et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 10012–10022
22. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (2017)



Towards Efficient Wound Management: An Automatic Deep Learning Model for Accurate Wound Image Segmentation

Aleksandar Miladinović¹✉, Alessandro Biscontin¹, Andrea Bonini²,
Francesco Bassi², Simone Kresevic², Alessandra Raffini², Katerina Iskra²,
Agostino Accardo², and Miloš Ajčević²

¹ Institute for Maternal and Child Health IRCCS “Burlo Garofolo”, Trieste, Italy
aleksandar.miladinovic@burlo.trieste.it

² Department of Engineering and Architecture, University of Trieste, Trieste, Italy

Abstract. Untreated wounds may lead to serious complications, underscoring the need for consistent monitoring. In clinical practice, wound imaging offers a potential tool for objective assessment. However, obtaining quantitative analysis requires accurate semantic image segmentation, a task that is often time-consuming. Despite deep learning has shown effectiveness in image segmentation, several challenges remain, especially in occluded or low-light images. This study aims to provide a novel semantic segmentation approach for wound images, designed to overcome some of these limitations. A semantic segmentation framework that combines a Convolutional Neural Network-based encoder with a lightweight multilayer perceptron decoder was proposed. A dataset consisting of 451 images was employed for the training, validation, and testing of a semantic segmentation architecture. During the training phase, the model achieved a pixel accuracy of 99.02%, with 98.82% on the validation set and 98.27% on the test set. The corresponding F1 scores were 0.97 for both training and validation, and 0.95 for the test set. The designed framework and the produced model provided a valuable solution for automatic wound assessment through precise image segmentation approach. The proposed method offers a remarkable contribution to telemedicine, providing a significant advancement for automatic wound management.

Keywords: Wound healing monitoring · Semantic segmentation · Deep Learning

1 Introduction

Chronic wounds significantly diminish patients’ quality of life [1]. If left untreated, they can result in severe complications, including limb amputation and even death [2]. Regular monitoring and appropriate treatment are crucial in preventing such deterioration. Besides the treatment, effective wound management involves continuous assessment, which is fundamental to track healing progress and determine the best treatment approach. Typically, this assessment relies on visual inspection of wound characteristics,

such as size, depth, and tissue growth. To enhance assessment and decision-making process, clinicians have integrated wound imaging using cameras and smartphones into wound evaluation protocol. These images are acquired during each examination, allowing professionals to track healing progress by comparing them with previous images. However, while wound imaging provides a more objective tool, it does not fully eliminate subjectivity in documentation. In fact, it may introduce additional variability due to the non-quantitative nature of visual comparisons and to external factors.

In recent years, the field of medical image analysis has experienced significant advancements, also in the field of wound assessment. Accurate measurement and segmentation of wound images play a pivotal role in diagnosing, treatments and monitoring healing processes. The complexities inherent in skin ulcer segmentation arise from the diverse characteristics of wounds, including variable shapes, sizes, and surrounding skin conditions [3]. Conventional techniques such as manual segmentation and simple thresholding often fail to meet the accuracy requirements for effective clinical applications, leading to a growing interest in automated, deep learning-based approaches [4]. Various methods have been employed for image segmentation in dermatological applications, predominantly grounded in traditional image processing techniques and machine learning paradigms [5, 6]. Methods such as color-based segmentation, contour detection, and superpixel algorithms have shown some efficacy in isolating wounds from background interference [7]. However, these methods often struggle with generalization across varied datasets and fail to perform well under adverse image conditions, such as non-homogeneous backgrounds or complex lighting [8, 9]. Additionally, many existing models focus more on retrieval and classification tasks rather than delivering the precision required for accurate segmentation [10]. The advent of deep learning has revolutionized the landscape of medical segmentation tasks; architectures like Fully Convolutional Networks (FCNs) and U-Net have gained prominence for their ability to capture intricate spatial relationships and learn semantic mappings between input images and their respective segmentation masks [11]. This has increased segmentation accuracy in challenging environments, establishing a benchmark for wound detection [4]. Despite these advancements, several challenges remain, including dependence on large, well-annotated datasets, susceptibility to misclassification in occluded or low-light conditions, and poor inter-observer agreement [4]. To address this, our approach involves refining a semantic segmentation architecture exploiting Contrast Limited Adaptive Histogram Equalization (CLAHE) as a preprocessing step. CLAHE is instrumental in enhancing image contrast, particularly in medical images, facilitating improved visibility of relevant structures in wound images that might otherwise be obscured [12]. This preprocessing technique enables the network to focus better on the wound areas, thereby improving segmentation outcomes [4]. The integration of CLAHE is expected to mitigate challenges related to illumination variations and enhance the overall robustness of the segmentation model.

The aim of this study is to develop an automated model for accurate wound segmentation, capable of effectively distinguishing wounds and linear scale markers from background elements. Through this work, we seek to provide a viable method that not only enhances accessibility and efficiency in wound assessment but also sets a foundation for future explorations of deep learning applications in clinical settings.

2 Materials and Methods

2.1 Dataset Definition

A dataset consisting of 451 images was employed for the training, validation, and testing of a semantic segmentation architecture. Additionally, an external test set comprising 98 images of the same category was utilized to further evaluate the model. The images were initially captured using smartphone cameras or similar devices and subsequently cropped to dimensions of 512x512 pixels. For each image, a corresponding ground truth mask was meticulously created. These generated masks highlighted two categories of regions of interest: “wound” and “scale marker.” Furthermore, a third class, designated as “background,” was established to include all other non-essential objects present within the image. The respective RGB triplets assigned to the wound, scale marker, and background regions were (0, 0, 0), (255, 255, 255), and (127, 127, 127) respectively.

2.2 Preprocessing

Prior to feeding the images into the network, a preprocessing stage was applied. First, noise reduction was performed applying a bilateral filter with a sigma spatial of 15, a sigma color of 15 and a filter size of 3px. To improve contrast, Contrast Limited Adaptive Histogram Equalization (CLAHE) was employed. The images were initially transformed from the RGB color space to the LAB color vector. To accomplish this, the RGB vector (normalized in the range [0,1]) was first linearized by applying the inverse of the standard gamma correction:

$$R_{lin} = \begin{cases} \frac{R}{12.92}, & \text{if } R \leq 0.04045 \\ \left(\frac{R+0.055}{1.055} \right)^{2.4}, & \text{if } R > 0.04045 \end{cases} \quad (1)$$

$$G_{lin} = \begin{cases} \frac{G}{12.92}, & \text{if } G \leq 0.04045 \\ \left(\frac{G+0.055}{1.055} \right)^{2.4}, & \text{if } G > 0.04045 \end{cases} \quad (2)$$

$$B_{lin} = \begin{cases} \frac{B}{12.92}, & \text{if } B \leq 0.04045 \\ \left(\frac{B+0.055}{1.055} \right)^{2.4}, & \text{if } B > 0.04045 \end{cases} \quad (3)$$

where R, G and B are the value of the red, green and blue channels of the image.

Then those vectors were multiplied by the transformation matrix:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \begin{bmatrix} R_{lin} \\ G_{lin} \\ B_{lin} \end{bmatrix} \quad (4)$$

Subsequently, to pass from XYZ space to the CIE LAB space (with the standard illuminant D65) the following transformations are applied:

$$L = 116f\left(\frac{Y}{Y_n}\right) - 16 \quad (5)$$

$$a = 500 \left[f \left(\frac{X}{X_n} \right) - f \left(\frac{Y}{Y_n} \right) \right] \quad (6)$$

$$b = 200 \left[f \left(\frac{Y}{Y_n} \right) - f \left(\frac{Z}{Z_n} \right) \right] \quad (7)$$

where X_n , Y_n , Z_n are the reference values for the D65 illuminant and

$$f(t) = \begin{cases} t^{1/3}, & \text{if } t > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3} \left(\frac{29}{6}\right)^2 t + \frac{4}{29}, & \text{otherwise} \end{cases} \quad (8)$$

Finally, the CLAHE algorithm was applied to the only perceptual lightness channel L of the images. The window size to implement the local adaptation was $32 \times 32 \text{ px}^2$ and the maximum slope of the probability density function of the resulting image was clipped to 0.5. Results of the preprocessing stage are shown in Fig. 1.

2.3 Semantic Segmentation Network

A semantic segmentation framework was explored for wound image analysis, integrating a Convolutional Neural Network (CNN)-based encoder with a lightweight multilayer perceptron decoder. The study focused on optimizing performance by varying the CNN backbone, replacing the encoder with multiple ResNet variants pre-trained on the ImageNet dataset. Only the best-performing model on the internal test set was retained for further evaluation. The output layer of the network was designed to match the dimensions of the input tensor, and a softmax activation function was applied in a depth-wise manner to generate a probability distribution for each pixel within the mask. Subsequently, the argmax function was applied to determine the predicted class for each pixel, categorizing it as either “wound,” “scale marker,” or “background.”

2.4 Training, Validation and Testing

Eighty images from the dataset were designated for internal testing, while the remaining three hundred and seventy-one images were allocated into training and validation sets, with 80% assigned to training and 20% to validation. The external test set was employed for additional metric evaluation. Data augmentation techniques were applied to the training set, allowing for transformations such as vertical flips, horizontal flips, and random rotations (specifically at angles of 90, 180, or 270°). The batch size was configured to 8, with the maximum number of epochs set at 500. A weighted categorical focal loss was utilized as the loss function, with the gamma value set to 2.5. Weights for the loss were assigned as follows: 0.45 for the “wound” category, 0.45 for the “scale marker” category, and 0.1 for the “background” category. Adam optimizer with an initial learning rate of 5×10^{-4} and a weight decay of 1×10^{-6} was utilized. To enhance training efficiency, early stopping was implemented, which halted the training phase if the validation loss failed to improve for more than eight epochs. Additionally, a learning rate scheduler was incorporated, whereby the learning rate was reduced to 0.25 times its value if the validation loss did not show improvement over a span of five epochs. The model that exhibited the lowest validation loss during the training phase was saved for subsequent evaluation on both the internal and external test sets.

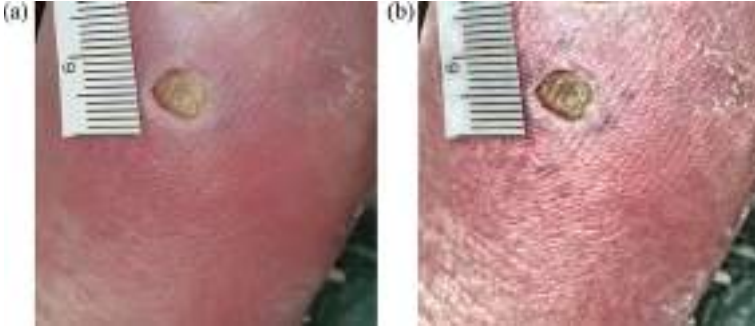


Fig. 1. Wound image before (a) and after (b) the preprocessing

2.5 Performance Metrics

To evaluate the performance of the proposed model, accuracy and F1 score were computed at both the training and evaluation stages. While the accuracy of an image was assessed at the pixel-level, the F1 score was first calculated for each class and then the total score for the image (F1_i) was obtained with the weighted average of the wound and the scale marker classes:

$$F1_i = \frac{3}{4}F1_w + \frac{1}{4}F1_s \quad (9)$$

where F1_w and F1_s are the F1 scores for wound and scale marker classes, respectively. To assess the overall performance of the fine-tuned model, an average F1 score was also computed as the average of the F1 scores:

$$F1_i = \frac{1}{n} \sum_{i=1}^n F1_i \quad (10)$$

3 Results

The best-performing model employed ResNet152 as the CNN backbone in the encoder. Table 1 presents the performance metrics (including accuracy, F1 score, and loss) across the training, validation, internal test, and external test sets for the best finetuned model. The model reached a pixel accuracy of 99.02% during training, while achieving 98.82% on the validation set and 98.27% on the test set. The corresponding F1 scores were 0.97 for both the training and validation sets, and 0.95 for the test set. The loss values were 7×10^{-4} for training, 1.2×10^{-3} for validation, and 2.2×10^{-3} for the test set. Finally, for this multi-class image segmentation problem, the weighted F1 scores were 0.87 and 0.85 on the internal and external test sets, respectively.

The result of segmentation using the proposed model, along with a comparison to the ground truth of two exemplary cases, is shown in Fig. 2.

Table 1. Accuracy, F1 Score, and Loss for Training, Validation, and Test Sets

Metric	Training	Validation	Internal test set	External test set
Accuracy (%)	99.02	98.82	98.27	–
F1 Score	0.97	0.96	0.96	–
F1 Score (Weighted)	–	–	0.87	0.88
Loss	0.7×10^{-3}	1.2×10^{-3}	2.2×10^{-3}	–

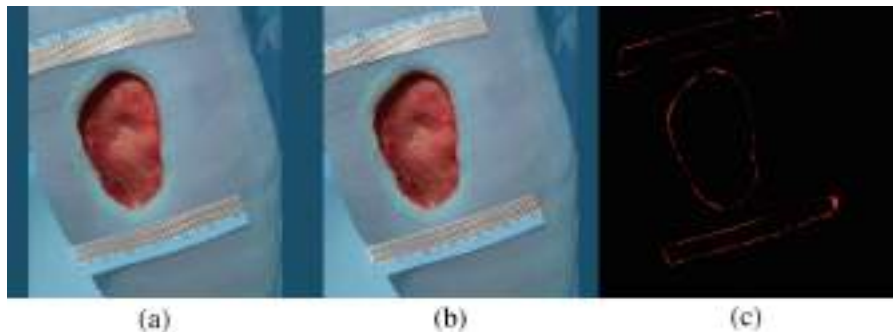


Fig. 2. Example of wound segmentation results produced by our model. (a) Segmentation output generated by our model, (b) ground truth annotation, and (c) visualization of differences between the model prediction and the ground truth, highlighting the detected discrepancies in red.

4 Discussion

Chronic wounds have a profound impact on patients’ quality of life and, if not properly managed, can lead to severe complications such as limb amputation or even mortality. Effective wound management extends beyond treatment alone; it requires continuous and quantitative monitoring and assessment, to track healing progression and optimize therapeutic strategies. The necessity of automatic evaluation tools is highlighted by the lack of inter-observer agreement and precision [13].

The field of medical image segmentation has witnessed rapid advancements in recent years, primarily due to the evolution of deep learning techniques. In particular, techniques based on encoder-decoder architectures have become gold standards in medical imaging due to their architecture’s ability to learn hierarchical features from complex datasets [14]. These techniques could be efficiently applied in the wound assessment field. Despite these advancements, several issues persist, such as dependence on large, annotated datasets and complexity of medical images themselves. Moreover, clinical applicability of segmentation algorithms depends on their robustness across diverse datasets. Variability in image acquisition modalities and patient demographics poses significant challenges, emphasizing the critical need for models to be both generalizable and interpretable [16].

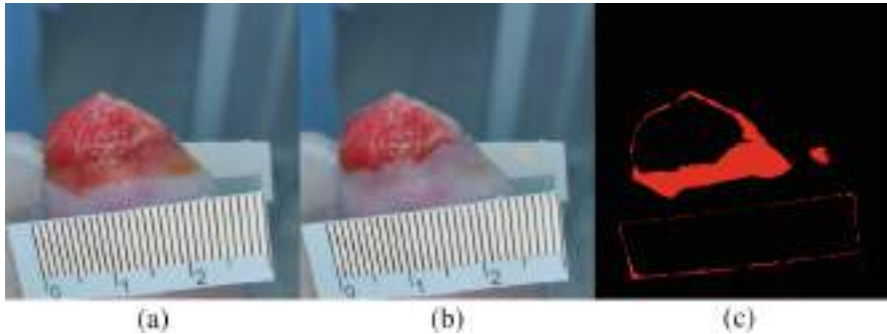


Fig. 3. Example of a segmentation failure due to the specific characteristics of the wound. (a) ground truth annotation; (b) predicted segmentation mask; (c) visualization of differences between the model prediction and the ground truth.

To address these challenges and further improve segmentation performance in complex medical imaging scenarios, this work provides novel deep learning architectures along with a specifically designed pre-processing step for medical image segmentation. By doing so, it aims to enhance both the effectiveness and functionality of medical image segmentation, with a particular focus on anatomically complex and difficult-to-detect structures, such as chronic wounds. The methodology involves the usage of an encoder-decoder architecture tailored for images derived from diverse medical modalities, alongside with custom preprocessing pipeline. Our model proved capable of detecting wounds with notable precision in most cases, although minor issues persist when the differentiation between damaged and healthy tissue is not clearly identifiable from the initial image (Fig. 3).

Our model showed remarkable performance with an accuracy of 0.98 and a F1-weighted of 0.87. Moreover, our results demonstrate strong generalization, as indicated by minimal performance degradation between the training and test sets.

The application of our segmentation technique could significantly support clinicians in assessing wound structures, providing a fast and reliable tool to make up for time-consuming processes and detection problems related to external factors. Moreover, considering that the healing of wounds is normally a long-lasting process, an evaluation method capable of providing accurate performances with smartphone-acquired camera is necessary to enable patients to stay at home and be frequently monitored at the same time [6]. This would lead to the applicability of our wound evaluation tool in telemedicine.

In conclusion, in this study by developing an automatic deep learning model for accurate wound image segmentation that demonstrates clinically relevant accuracy, contributes to the ongoing evolution of advanced wound assessment and management and lays the groundwork for future clinical applicability.

References

1. Lindholm C, Searle R (2016) Wound management for the 21st century: combining effectiveness and efficiency. *Int Wound J* 13:5–15. <https://doi.org/10.1111/iwj.12623>

2. Sen CK (2019) Human wounds and its burden: an updated compendium of estimates. *Adv Wound Care* (New Rochelle) 8:39–48. <https://doi.org/10.1089/wound.2019.0946>
3. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43:17–29. <https://doi.org/10.1016/j.bbe.2022.11.004>
4. Alabdulhafith M, Ba Mahel AS, Samee NA et al (2024) Automated wound care by employing a reliable U-Net architecture combined with ResNet feature encoders for monitoring chronic wounds. *Front Med* (Lausanne) 11:36977. <https://doi.org/10.3389/fmed.2024.1310137>
5. Chino DYT, Scabora LC, Cazzolato MT et al (2020) Segmenting skin ulcers and measuring the wound area using deep convolutional networks. *Comput Methods Programs Biomed* 191:105376. <https://doi.org/10.1016/j.cmpb.2020.105376>
6. Foltynski P, Ciechanowska A, Ladyzynski P (2021) Wound surface area measurement methods. *Biocybern Biomed Eng* 41:1454–1465. <https://doi.org/10.1016/j.bbe.2021.04.011>
7. Blanco G, Traina AJM, Traina C et al (2020) A superpixel-driven deep learning approach for the analysis of dermatological wounds. *Comput Methods Programs Biomed* 183:105079. <https://doi.org/10.1016/j.cmpb.2019.105079>
8. Hsu JT, Chen YW, Ho TW, et al (2019) Chronic wound assessment and infection detection method. *BMC Med Inform Decis Mak* 19. <https://doi.org/10.1186/s12911-019-0813-0>
9. Wang J, Zhang H, Yi Z (2023) CCTrans: improving medical image segmentation with contoured convolutional transformer network. *Mathematics* 11. <https://doi.org/10.3390/math11092082>
10. Zhang R, Tian D, Xu D et al (2022) A Survey of Wound image analysis using deep learning: classification, detection, and segmentation. *IEEE Access* 10:79502–79515. <https://doi.org/10.1109/ACCESS.2022.3194529>
11. Ramachandram D, Ramirez-GarciaLuna JL, Fraser RDJ et al (2022) Fully automated wound tissue segmentation using deep learning on mobile devices: cohort study. *JMIR Mhealth Uhealth* 10:36977. <https://doi.org/10.2196/36977>
12. Marcato B, Ferraz C, Verri W, et al Automatic area estimation of mice wound images, 1–8. <https://doi.org/10.5753/kdmile.2024.241973>
13. Kekonen A, Bergelin M, Eriksson J-E et al (2017) Bioimpedance measurement based evaluation of wound healing. *Physiol Meas* 38:1373–1383. <https://doi.org/10.1088/1361-6579/aa63d6>
14. Siddique N, Paheding S, Elkin CP, Devabhaktuni V (2021) U-Net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9:82031–82057. <https://doi.org/10.1109/ACCESS.2021.3086020>
15. Antonelli M, Reinke A, Bakas S, et al (2022) The Medical Segmentation Decathlon. *Nat Commun* 13. <https://doi.org/10.1038/s41467-022-30695-9>
16. Tajbakhsh N, Jeyaseelan L, Li Q et al (2020) Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal* 63:101693. <https://doi.org/10.1016/j.media.2020.101693>



Deep Learning-Based Wound Segmentation Using a U-Net Architecture

Martinus Boom^(✉)

Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
uc2024154464@student.uc.pt

Abstract. Inspired by the IFMBE Scientific Challenge Competition at NBC 2025, this work presents a deep learning approach for semantic wound segmentation employing a U-Net architecture. The approach focuses on segmenting three components within wound images: wound area, background, and scale marker. Preprocessing included grayscale conversion, resizing, and data augmentation on a dataset consisting of 371 training images and 80 validation images, each with associated 512×512 pixel masks. On the validation set, the model gets a weighted F1 score of 0.8327 using an encoder-decoder structure with skip connections. Post-processing Lanczos interpolation guarantees that output masks satisfy the criteria of the challenge. The best model achieved a test set F1 score of 0.7924, as reported by the challenge organizers.

Keywords: Wound Segmentation · U-Net · Deep Learning · Medical Imaging · Data Augmentation · Image Resizing

1 Introduction

In medical image analysis, wound segmentation is an important chore as it helps to monitor, evaluate, and plan the treatment of wounds. Often depending on manual delineation or basic thresholding techniques, traditional segmentation methods are labor-intensive and subject to observer variability. Deep learning techniques have lately become a useful substitute by automatically learning hierarchical representations from image data.

One well-known example is the U-Net architecture [1], which efficiently integrates contextual and spatial information through an encoder-decoder structure with skip links. Proposed to solve different difficulties in biomedical image segmentation are extensions of this architecture. For semantic segmentation, Long et al. [2] presented fully convolutional networks; çiçek et al. [3] modified U-Net for volumetric data. Lately, U-Net variants have included attention mechanisms [4] to enhance segmentation performance even more.

This work applies a U-Net-based method to the task of wound segmentation (see Fig. 1), distinguishing between wound areas, background, and scale markers. The study utilizes a dataset of training and validation images, where images are preprocessed, augmented, and resized for efficient model training. Although a

test dataset is defined in the IFMBE Scientific Challenge Competition guidelines at NBC 2025 [?], only training and validation data were available for this study. Reported challenge results include a final test F1 score of 0.7924 for one attempt and a validation F1 score of 0.8327.

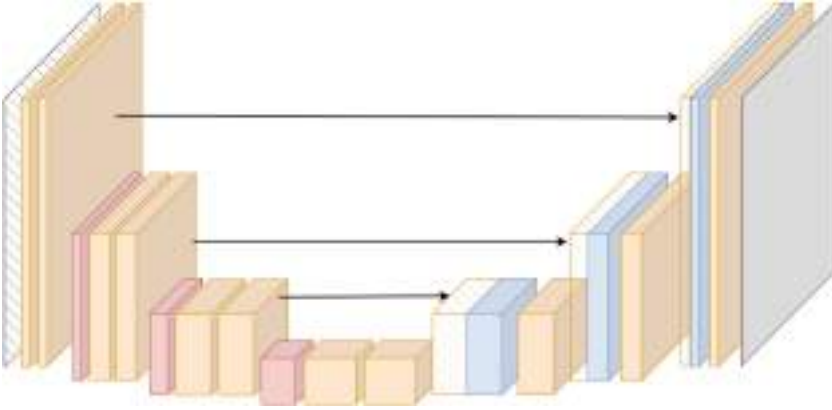


Fig. 1. U-Net Architecture [6].

The dataset used in this work was originally introduced in [5] as part of the AutoPlanimator Project [7], although only training and validation datasets were directly available for the present study.

2 Related Work

Medical image segmentation has seen rapid advances with the emergence of deep learning. Among these, U-Net [1] has become a reference architecture due to its simplicity and strong performance in biomedical tasks. Originally proposed for cell segmentation, U-Net uses an encoder-decoder structure with skip connections to combine semantic and spatial information—a balance that is essential for tasks like wound segmentation.

Several extensions have since been proposed. çİçek et al. [8] introduced 3D U-Net for volumetric data, through applications in MRI and CT scans. After, attention mechanisms were incorporated to highlight relevant features in the image. For example, Oktay et al. [9] proposed the Attention U-Net, improving segmentation performance by focusing on informative regions.

Currently, transformer-based architectures such as TransUNet [10] have demonstrated competitiveness for the segmentation of medical images in modeling long-range dependencies. However, typically, such models would require larger datasets and compute resources, and in some scenarios, it would not be possible—especially where one has very small domain-specific datasets such as the situation presented here.

In the context of wound segmentation specifically, Foltynski and Ladyzynski [11] proposed a deep learning framework to measure wound areas using adaptive calibration and U-Net-based segmentation. Their work demonstrated that accurate segmentation of wound regions and scale markers is achievable and clinically valuable.

Here, the choice of U-Net achieves a good balance between performance, simplicity, and computational cost. Since more complex architectures can deliver incremental gains, U-Net is still a good baseline to beat—particularly under low-data and low-resource conditions.

3 Methodology

The segmentation of the wound included data preparation, neural network architectural development, training, and the final postprocessing of the scale. Each of those steps was carefully developed to maximize segmentation accuracy and meet the challenge’s requirements. The data preparation phase included scouring for training images and 86 validation images in JPG format, starting from a resolution of 512×512 pixels. Each of the images had a PNG ground truth mask (see Fig. 2) in which scale markers were indicated in white (RGB: 255,255,255), the background areas in gray (RGB: 127,127,127), and the areas of the wound in black (RGB: 0,0,0). Every image had one or more indications of scale incorporated into it and one or more areas of the wound. To help with learning, these images were converted to single-channel greyscale and had the resizing done later to 256×256 pixels. Acknowledging that variability is extremely important when it comes to the robustness of any model, applied data augmentation techniques, including random rotation and horizontal and vertical flipping.



Fig. 2. Example of image with wound and scale markers corresponding ground truth mask.

The U-Net-based segmentation model is the well-known U-Net architecture. The encoder part of the U-Net had a number of sequential double convolution

blocks, with downsampling carried out after each block. The decoder of this U-Net follows the encoder in its architecture, employing transposed convolutions for upsampling. Also made use of skip connections between the encoder and decoder to retain certain lost fine (more detailed) information present in the original image after the downsampling step. All of the convolution blocks had Batch Normalization and made use of the ReLU activation function to enhance the efficiency of training. The Cross-Entropy Loss was the choice for the training process as it extracts segmentation accuracy at a pixel level efficiently. The Adam optimizer was used for model parameter optimization, while the learning rate and batch size were determined by numerous trials. Segmentation accuracy on the validation dataset was used to closely monitor the progress of model training and performance.

To perform a quantitative evaluation of model performance, utilized a weighted F1 score based on the performance of the model on wound segmentation while taking into consideration scale marker segmentation. The computed F1 scores indicated wound pixels (black) and scale markers (white) independently, and then added these two into a weighted F1 score, given as follows:

$$F1_i = \frac{3}{4} F1_{\text{wound}} + \frac{1}{4} F1_{\text{scale}}$$

Post-processing and Interpolation Method

After segmentation, the predicted masks (originally sized at 256×256) needed to be resized back to 512×512 pixels to comply with the challenge submission requirements. For this task, we selected the Lanczos interpolation method. Lanczos is a high-quality resampling filter that preserves edge sharpness and spatial detail better than simpler alternatives.

Lanczos was compared with bilinear and nearest-neighbor interpolation. Nearest-neighbor is fast but produces blocky output and deforms fine boundaries, which is highly critical in medical segmentation tasks. Bilinear interpolation smooths boundaries but deforms significant edges, especially in masks where class boundaries are critical. Lanczos provided a good balance—preserving the integrity of the segmented regions without over-smoothing. Empirically, was observe that Lanczos application provided class separability between the masks and less post-processing artifact.

4 Results

The last model was validated on a validation set that yielded an F1 score of 0.8327, reflecting the combined performance of wound segmentation and scale marker. Statistical information was not collected during the experiments, so no additional curves, nor quantitative data are available to report the progress of loss or accuracy per epoch. Although not directly accessible by the participants, the test dataset outcomes of 3 attempts are reported by the challenge organizers in Table 1. The first false one was eliminated for wrong output size, the second one scored an F1 of 0.7329, and the third one scored an F1 of 0.7924. These

performances denote a certain potential in regards to the applicability of the model towards unseen data, even in the absence of its direct testing in this study.

Table 1. Outcomes of Three Attempts on the Challenge Test Dataset (as Reported by Organizers)

Attempt	Outcome	F1 Score
1	Rejected (incorrect mask size)	–
2	Accepted	0.7329
3	Accepted	0.7924

4.1 Comparative Analysis

To better understand the effectiveness of U-Net-based segmentation model, it was compared with a simple baseline method, the Otsu thresholding. While this method is more common used in traditional image processing, it fails on the ability to distinguish between complex structures like wounds and scale markers in real-world medical images.

In the test phase, Otsu thresholding applied to grayscale versions of the images resulted in noisy segmentation, with frequent misclassifications of the background as wound area. This approach results in an average weighted F1 score of approximately 0.5476, far below the 0.8327 achieved by our U-Net model on the validation set.

This comparison highlights the added value of deep learning in handling the variability and complexity present in medical images, especially in tasks requiring both accuracy and consistency.

5 Discussion

On account of the analysis of the validation dataset, it was observed with efficacy that the U-Net architecture is effective in segmenting the wounds and the scale markers once the algorithm trained itself on grayscale images of size 256×256 pixels. Data augmentations such as random flips and random rotations seem to contribute enormously to the general robustness of the model, with an F1 score being computed at 0.8327. The gap observed, however, in the validation (F1 score: 0.8327) and the testing performance (the highest reported F1 score: 0.7924), implied some overfitting to training and validation datasets or might reflect some intrinsic differences in data distributions between the validation and test datasets. [12].

The performance on test data that is unseen, according to the challenge organizers, signifies the challenges on segmentation models face when being generalized on a reduced dataset. Previous work has reported the same observations,

adding that complete training data as well as sophisticated augmentation techniques are necessary to boost generalization.[13].

Training on the original resolution of 512×512 pixels may be an upgrade in the future to restrict the scope for interpolation antiques to reduce segmentation accuracy [14]. Application of more sophisticated variants of U-Net, e.g., attention to U-Net architectures [15], could also improve performance by highlighting relevant spatial detail. Increasing the connectivity of layers using deeper convolutional architectures or transformer-based architecture could be considered to further optimize accuracy and generalization [16]. To provide a more accurate description of model performance, future evaluations must also include other segmentation measures like the Dice coefficient or Intersection over Union (IoU) [17].

6 Limitations

While the proposed approach showed promising results, there are several limitations that should be considered.

First, although the model achieved a strong F1 score of 0.8327 on the validation set, the performance on the test set dropped to 0.7924. This gap may be attributed to overfitting the training/validation data and possible distributional differences between the test set. Without access to the test data for independent analysis, it is difficult to identify where is the difference across from domain shifts (e.g., lighting, wound type, background variation).

Second, to facilitate computational limitations, all training and validation images were downscaled from 512×512 to 256×256 pixels. Although this allowed faster training and reduced memory usage, it likely led to the loss of fine-grained spatial details that are critical in tasks like wound edge detection or small-scale marker segmentation. The resizing may also have contributed to interpolation artifacts, especially in boundary regions between classes.

Furthermore, hyperparameter tuning was performed manually due to resource constraints. A more systematic approach (e.g. grid search) could have found better configurations for the parameters.

Finally, the evaluation was based on a weighted F1 score. Although appropriate for the challenge, complementary metrics such as Dice coefficient, Intersection over Union (IoU), and sensitivity-specificity analyzes would provide a broader perspective on the model performance and robustness.

Future work should explore training directly on full-resolution images, introduce more sophisticated architectures (e.g., attention U-Nets or transformer-based models), and implement cross-validation or external validation sets to improve generalization and reduce overfitting.

7 Conclusion and Future Work

For this research was developed a deep-learning-based wound segmentation using a U-Net architecture focusing on delineation of wound areas, background, and

scale markers. The fundamental stages of the deep learning-based wound segmentation technique using U-Net architecture involve data preprocessing, converting to grayscale, resizing it to lower dimensions of 256×256 pixels, and data augmentation techniques to enhance the generalization capabilities of the model. The U-Net model successfully learned the spatial features for accurate segmentation using an encoder-decoder architecture with skip connections, Batch Normalization, and ReLU activations. The model showed quite good performance on the validation set, recording a weighted F1 score of 0.8327; However, the results reported by the challenge organizers indicate that the model generalized reasonably well, achieving a test F1 score of 0.7924.

Further work in the setting of the original resolution of 512×512 pixels is doggedly necessary to reduce the interpolation impairment factor and ensure further robustness of the segmentation scheme, along with hyperparameter tuning and applying these frameworks, such as attention-based U-Net, deeper convolution-based techniques, or even transformer-based ones. Other evaluation metrics, besides those the Dice coefficient, Intersection over Union (IoU), and sensitivity-specificity analysis allow for a clearer and more complete picture of the model performance and confidence.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. *Med. Image Comput. Comput.-Assist. Interven. (MICCAI)* 234–241 (2015)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440 (2015)
3. Çiçek Ö., Abdulkadir A., Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W (eds) *MICCAI 2016*, vol 9901. LNCS. Springer, Cham, pp 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
4. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: a nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11 (2018)
5. IFMBE scientific challenge competition 2025: segmentation challenge: wounds and markers in focus. <https://nbc2025.ibib.waw.pl/index.php/ifmbe-scientific-challenge-competition/>
6. O'Sullivan, C.: U-Net explained: understanding its image segmentation architecture, *medium.com*
7. AutoPlanimator Project. <https://autoplanimator.ibib.waw.pl/>
8. çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. *Med. Image Comput. Comput.-Assist. Interven. (MICCAI)* 424–432 (2016)
9. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. *Med. Imaging Deep Learn. (MIDL)* 1–10 (2018)
10. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021)

11. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern. Biomed. Eng.* 43:17–29. <https://doi.org/10.1016/j.bbe.2022.11.004>
12. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016)
13. Isensee, F., et al.: nnU-Net: self-adapting framework for u-net-based medical image segmentation. *arXiv preprint [arXiv:1809.10486](https://arxiv.org/abs/1809.10486)* (2019)
14. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* **109**(5), 820–838 (2020)
15. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. In: *Medical Imaging with Deep Learning (MIDL)*, pp. 1–10 (2018)
16. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)* (2020)
17. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15(1):1–28



Dual-Attention U-Net++ with Class-Specific Ensembles and Bayesian Hyperparameter Optimization for Precise Wound and Scale Marker Segmentation

Daniel Cieślak^{1,2(✉)}, Miriam Reca¹, Olena Onyshchenko¹,
and Jacek Rumiński¹

¹ Faculty of Electronics, Telecommunications and Informatics,
Gdańsk University of Technology, Gdańsk, Poland
cieslak.a.daniel@gmail.com

² IDEAS NCBR, Warsaw, Poland

Abstract. Accurate segmentation of wounds and scale markers in clinical images remains a significant challenge, crucial for effective wound management and automated assessment. In this study, we propose a novel dual-attention U-Net++ architecture, integrating channel-wise (SCSE) and spatial attention mechanisms to address severe class imbalance and variability in medical images effectively. Initially, extensive benchmarking across diverse architectures and encoders via 5-fold cross-validation identified EfficientNet-B7 as the optimal encoder backbone. Subsequently, we independently trained two class-specific models with tailored preprocessing, extensive data augmentation, and Bayesian hyperparameter tuning (WandB sweeps). The final model ensemble utilized Test Time Augmentation to further enhance prediction reliability. Our approach was evaluated on a benchmark dataset from the NBC 2025 & PCBBE 2025 competition. Segmentation performance was quantified using a weighted F1-score (75% wounds, 25% scale markers), calculated externally by competition organizers on undisclosed hardware. The proposed approach achieved an F1-score of 0.8640, underscoring its effectiveness for complex medical segmentation tasks.

Keywords: Image segmentation · Wound assessment · Deep learning · Medical imaging · U-Net++ · Attention Mechanism · Bayesian Optimization

1 Introduction

The skin, being a vital organ, is usually categorized as the body's largest. The complex, layered structure acts as a safeguard. Although strong, it is simultaneously sensitive and easily affected by injuries from accidents, extended illnesses,

surgeries, and other causes. Proper and precise documentation of wounds is essential for their effective healing. In addition, detecting skin conditions accurately is vital for reducing incidence, treatment costs, morbidity, and mortality, highlighting the need to examine damaged skin [1]. Medical practitioners have historically relied on manual measurements, photos, and visual assessments to track and evaluate wound healing. While these techniques are generally effective, they are subjective, prone to mistakes, and depend on the healthcare provider’s experience [2]. This variability emphasizes the pressing need for technological progress in wound evaluation.

Recent advancements in computer vision and deep learning present a compelling alternative for manual methods. Automated segmentation methods promise objectivity, consistency, and efficiency by transforming pixels on a medical image into insightful clinical data. However, developing an effective segmentation system isn’t straightforward; wounds can be irregularly shaped, vary significantly in appearance, and imaging conditions may differ greatly. Additionally, precise measurements require reliable scale markers within images, further complicating the task.

Our goal for the NBC 2025 & PCBBE 2025 wound image segmentation challenge was not only to benchmark cutting-edge neural networks but also to contribute to a broader narrative—shifting from manual subjectivity toward accurate, reproducible wound assessments powered by artificial intelligence.

In this paper, we present our approach—from thorough exploration of contemporary deep-learning architectures, through rigorous experimentation, to identifying the most promising strategy for simultaneously segmenting wounds and scale markers. Our findings illustrate the strengths and limitations of current state-of-the-art methods and set the stage for further innovations, bringing us closer to making automated, reliable wound assessments a clinical reality.

2 Related Work

Classical image processing techniques, such as thresholding, contour models and region segmentation, rely on handcrafted features and minimal labeled data but face challenges with complex textures and lighting variations [3]. The U-Net architecture [4] was initially introduced for medical image segmentation, its main advantage is that it can be trained on limited-size datasets. A dual attention U-Net model was proposed for wound segmentation, integrating VGG16 and U-Net architectures with dual attention mechanisms to focus on relevant regions within the wound area. This model achieved high performance metrics, including a Dice coefficient of 94.1% and IoU of 89.3% on test data [5] (Table 1).

Class imbalance in wound segmentation datasets leads to biased model performance, with studies reporting mean IoU as low as 48.6% for minority classes like subcutaneous hematomas, but strategies like dual-encoder architectures (SRU-Net) and loss weighting improve accuracy in multi-class scenarios, achieving reliable classification for 7 common wound types despite imbalance [9]. The use of attention mechanisms in wound segmentation models could enhance their

Table 1. Comparison of different neural network architectures for segmentation tasks

Architecture	Strengths	Weaknesses	Multi-class performance
SegFormer [6]	Combines transformers with lightweight decoders. Captures global context and fine-grained details efficiently.	Requires larger datasets. May underperform on highly imbalanced classes.	Excellent for multi-class tasks with sufficient data.
U-Net++ [7]	Nested skip pathways and deep supervision improve accuracy for multi-scale structures. Reduces semantic gaps.	Increased model complexity. Higher computational costs.	Superior for multi-scale structures like wounds and markers.
DeepLab V3+ [8]	Captures multi-scale contextual information. Handles class imbalance well.	Computationally intensive. Less effective on small datasets.	Excels in multi-class scenarios with sufficient resources.
U-Net	Simple encoder-decoder structure. Reliable with limited data. Widely validated.	Limited contextual awareness. Struggles with fine-grained multi-class tasks.	Robust baseline but less effective for complex multi-class tasks.

ability to focus on relevant features, significantly improving accuracy by selectively weighting spatial and channel-wise information, as demonstrated in dual attention U-Nets achieving high Dice coefficients in medical image segmentation tasks [10].

While recent dual-attention architectures reported promising Dice coefficients (e.g., 94.1% by [5]), our study further improves segmentation accuracy through the strategic choice of a stronger encoder (EfficientNet-B7), rigorous Bayesian hyperparameter tuning, and dedicated class-specific ensembles, directly addressing the severe class imbalance and variability present in clinical images

3 Methods

3.1 Dataset

We utilized two distinct datasets to ensure robust training and evaluation of our segmentation models. The primary dataset, provided by the NBC 2025 & PCBBE 2025 challenge organizers, contained images annotated with wounds and scale markers.

Specifically, the training set comprised 371 images in total, of which 208 had two labels (annotated with wounds and scale markers) and 163 without second label. The validation set consisted of 80 images, with 79 annotated with 2 labels and only 1 labeled with one class, clearly highlighting the imbalance issue we aimed to address in our methodological approach.

The secondary dataset, sourced from MICCAI 2020 [11], included only wound annotations, serving as an auxiliary dataset for pretraining.

Due to significant class imbalance, we trained separate models for wounds and scale markers, each optimized individually with customized preprocessing.

3.2 Image Preprocessing

We optimized preprocessing strategies separately for each class, employing intensive data augmentations, including flips, rotations, brightness-contrast

adjustments, and geometric distortions. Images were normalized based on ImageNet statistics. Augmentation strength was optimized through Bayesian sweeps (WandB), effectively enhancing generalization.

Augmentations were applied to 100% of our training images using the Albumentations library. Specifically, these included horizontal and vertical flips (probability 50%), random rotations by multiples of 90° (probability 50%), random brightness adjustments (limits $[-0.2274, +0.4548]$) and contrast adjustments (± 0.2274 , probability 80%), geometric distortions such as grid distortion (limit ± 0.3411 , probability 50%), and grid dropout (ratio 0.3, probability 50%). The augmentation strength was optimized through Bayesian sweeps using WandB, with the final strength parameter set to 1.137.

3.3 Segmentation Architectures

Initially, extensive benchmarking was conducted using Segmentation Models PyTorch (SMP) library architectures: U-Net, U-Net++, MAnet, LinkNet, FPN, PSPNet, PAN, DeepLabV3, DeepLabV3+, UPerNet, and SegFormer. These architectures were evaluated across multiple encoders (DenseNet121, EfficientNet-B0, ResNet34, VGG16, MiT-B0, and MobileNetV2) using 5-fold cross-validation on NVIDIA A5000 GPUs. Benchmark results (Fig. 1, Fig. 2) identified U-Net++ with EfficientNet-B7 encoder as superior.

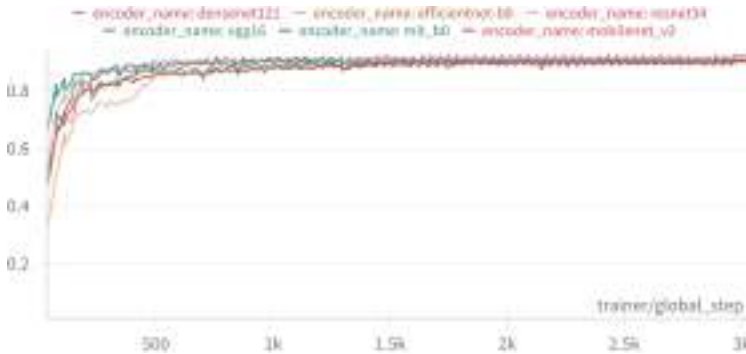


Fig. 1. Encoder backbone benchmarking results (validation Dice scores).

We further enhanced U-Net++ by integrating a dual-attention module: spatial attention and SCSE channel-wise attention.

3.4 Training Protocol

Models for wounds and markers were trained independently due to data imbalance.

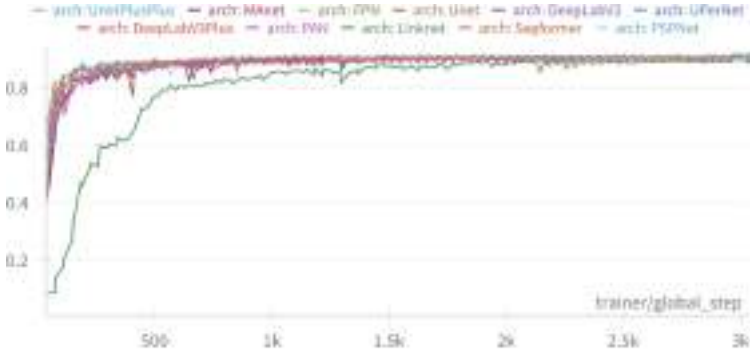


Fig. 2. Performance comparison of segmentation architectures (EfficientNet-B0 encoder).

This approach effectively mitigated class imbalance, enabling targeted optimization for each specific class. Separate training allowed easier prototyping and rapid introduction of corrective measures for particularly challenging segmentation cases, significantly improving the overall robustness and performance of the solution.

Both utilized dual-attention U-Net++ with EfficientNet-B7 encoder, optimized via Bayesian hyperparameter sweeps on WandB (loss ratios, augmentation strength, learning rate, mixup alpha).

We trained models using PyTorch Lightning, distributed across two NVIDIA A5000 GPUs, with mixed-precision training. Training ran for 80 epochs, monitored via validation Dice.

Test Time Augmentation aggregated predictions from rotations (0° , 90° , 180° , 270°). Conditional Random Fields were optionally applied post-inference. Final performance was averaged across five runs.

3.5 Implementation Details

Key hyperparameters (optimized via Bayesian hyperparameter sweeps using WandB) included:

- Architecture: U-Net++ (dual-attention: SCSE + Spatial)
- Encoder: EfficientNet-B7 (ImageNet pretrained)
- Optimizer: AdamW (lr=0.001; ReduceLROnPlateau scheduler)
- Loss ratios optimized via WandB
- Batch size: 4; Epochs: 80
- Mixup alpha: 0.3604; Augmentation strength: 1.137
- Precision: Mixed (FP16)

3.6 Overall Framework Structure

To clearly illustrate our approach, we present an overview of our segmentation pipeline in Fig. 3. The pipeline begins with preprocessing and intensive data augmentation, followed by the independent training of class-specific dual-attention

U-Net++ models with EfficientNet-B7 encoders. Finally, ensemble predictions are generated using Test Time Augmentation and optionally refined via Conditional Random Fields post-processing.

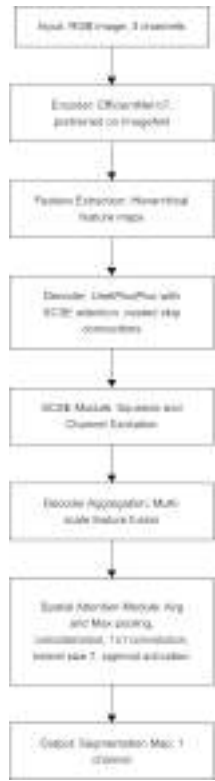


Fig. 3. Overview of the segmentation pipeline: preprocessing, independent training of dual-attention U-Net++ models, ensemble predictions with TTA, and optional CRF-based refinement.

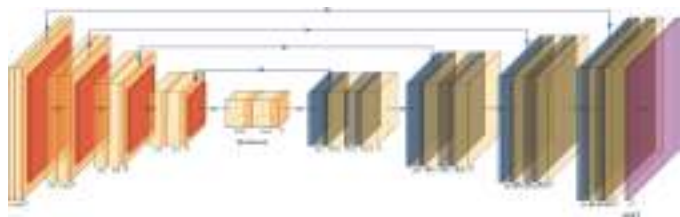


Fig. 4. Schematic visualization of the dual-attention U-Net++ architecture with EfficientNet-B7 encoder. Attention modules (Spatial + SCSE) recalibrate feature maps, emphasizing relevant spatial regions and discriminative channels.

3.7 Dual-Attention U-Net++ Architecture

A detailed visualization of the dual-attention U-Net++ architecture is provided in Fig. 4. The model uses EfficientNet-B7 as the encoder backbone to extract high-level image features. Nested skip-connections characteristic of U-Net++ facilitate efficient feature reuse. Dual-attention modules, combining spatial and SCSE (channel-wise) attention mechanisms, are integrated within the decoding pathway to enhance segmentation accuracy.

4 Results

Our proposed dual-attention U-Net++ architecture, optimized via Bayesian sweeps, was evaluated on the NBC 2025 & PCBBE 2025 segmentation challenge. Due to dataset imbalance, we trained separate models for wounds and scale markers, achieving robust internal validation scores as summarized in Table 2.

Table 2. Internal validation Dice scores for wounds and scale markers across four different training runs. Best results for each class are shown in bold.

Run	Wound segmentation (Dice)	Scale marker segmentation (Dice)
1	0.9693	0.9028
2	0.9743	0.9281
3	0.9733	0.9175
4	0.9705	0.9258
Average	0.9718	0.9185

Final submissions were based on an ensemble of two scale marker models combined with a single, robust wound segmentation model.

The primary evaluation metric, chosen by the organizers, was a weighted F1-score (75% wound segmentation, 25% scale marker segmentation), computed externally on undisclosed hardware to ensure fairness. At the time of manuscript preparation, our models’ official evaluation by competition organizers is ongoing; preliminary internal tests suggest high reliability and robustness.

Our final externally validated weighted F1-score was 0.8640.

5 Discussion

Segmenting wounds and scale markers poses significant challenges, including class imbalance, variable imaging conditions, and complex boundaries. Our approach addressed these issues through specialized training pipelines, dual-attention mechanisms, and Bayesian hyperparameter tuning.

Comparison with Similar Studies. Similar research efforts, such as those by Niri et al. [5] and Zimmermann et al. [10], explored attention-enhanced U-Net

architectures. Niri et al. reported impressive Dice scores (approximately 94.1%) with a dual-attention VGG16-based U-Net architecture. Our study extends these prior efforts by leveraging the powerful EfficientNet-B7 backbone, rigorous Bayesian hyperparameter optimization, and dedicated class-specific models. Unlike previous single-model multi-class approaches, our method facilitated targeted optimizations to handle severe class imbalance more effectively.

Clarification on EfficientNet Encoder Selection. While our initial benchmarking results (Fig. 1, Fig. 2) demonstrated EfficientNet-B0 as the top-performing encoder, we chose EfficientNet-B7 due to its architectural similarity yet greater representational capacity. Previous studies consistently highlight that higher-capacity EfficientNet variants significantly enhance generalization and robustness, which aligns with our objective of achieving superior segmentation performance in complex and highly variable medical imaging contexts.

Initially, extensive benchmarking identified EfficientNet-B7 as optimal despite its computational demands, justified by significant performance gains in validation Dice scores. The dual-attention enhancement of U-Net++ provided significant improvements in capturing spatial and channel-wise context.

Independent training for each class proved essential in managing dataset imbalance, aligning with contemporary research [11]. Furthermore, adaptive loss weighting strategies demonstrated clear advantages, although occasionally causing unstable training, including sudden learning rate spikes leading to training collapse (Dice dropping to 0), as visualized in Fig. 6.

Test Time Augmentation consistently improved segmentation accuracy, emphasizing its role in real-world medical imaging. Future work could explore transformer-based architectures and more sophisticated ensemble strategies to improve accuracy further.

6 Conclusion

In this study, we developed a robust and effective segmentation framework based on a dual-attention enhanced U-Net++ architecture for precise wound and scale marker segmentation. Our systematic benchmarking across various neural network architectures and encoder backbones identified EfficientNet-B7 as the optimal encoder, significantly enhancing segmentation accuracy and generalization capabilities. To address the pronounced class imbalance, we implemented class-specific training strategies, including tailored preprocessing, extensive data augmentation, and Bayesian hyperparameter optimization. These measures were critical in stabilizing model performance and achieving superior segmentation quality.

The incorporation of dual-attention mechanisms-combining spatial and channel-wise (SCSE) attention modules-proved essential for capturing complex features inherent in medical imagery, thereby improving both localization accuracy and boundary precision. Moreover, the use of ensemble prediction strategies combined with Test Time Augmentation further enhanced the reliability and robustness of our approach.

Evaluated externally on a challenging benchmark provided by the NBC 2025 & PCBBE 2025 competition, our proposed method achieved a weighted F1-score of 0.8640, demonstrating significant potential for practical, clinical adoption in wound management workflows. These results underscore the benefits of advanced architectural enhancements, rigorous hyperparameter tuning, and ensemble methodologies for tackling complex, imbalanced medical image segmentation tasks.

Future work will explore the integration of transformer-based models, more sophisticated post-processing techniques, and the evaluation of the framework on diverse clinical datasets to further improve generalization and reliability, bringing automated, accurate wound assessment closer to routine clinical practice.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** Authors declare no conflicts of interest.
- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** Dataset provided by NBC2025 Challenge organizers.
- **Materials availability:** Not applicable.
- **Code availability:** Available upon reasonable request.
- **Author contribution:** Conceptualization and methodology: DC, JR, MR, OO; Software, validation, original draft preparation: DC; Review & editing: MR, OO, JR.

Appendix A Additional Training Details

All models were implemented using PyTorch Lightning and trained on NVIDIA A5000 GPUs using mixed precision. The final ensemble of scale-marker segmentation models was implemented as follows (Fig. 5):

Algorithm 1. Ensemble Prediction with CRF Postprocessing

Require: Image I , Trained models M_1, M_2 , weights w_1, w_2 , CRF iterations

- 1: $\text{prob1} \leftarrow \text{sigmoid}(M_1(I))$
 - 2: $\text{prob2} \leftarrow \text{sigmoid}(M_2(I))$
 - 3: $\text{ensemble\_prob} \leftarrow w_1 \cdot \text{prob1} + w_2 \cdot \text{prob2}$
 - 4: $\text{binary\_mask} \leftarrow \text{ensemble\_prob} > 0.5$
 - 5: $\text{final\_mask} \leftarrow \text{CRF\_postprocess}(I, \text{binary\_mask}, \text{iterations})$
 - 6: **return** final_mask
-



Fig. 5. Training validation Dice scores for class 1 (wounds). Model demonstrated stable convergence around 97%.

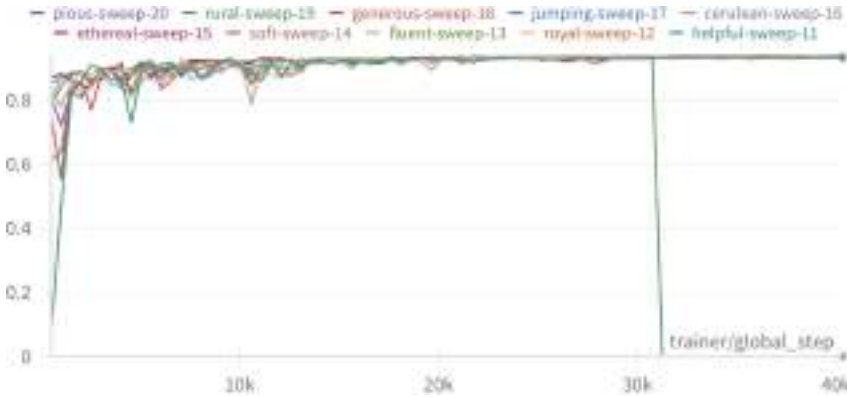


Fig. 6. Training validation Dice scores for class 2 (scale markers). Note the instability due to adaptive loss functions, including a sudden collapse (Dice ≈ 0) during training.

Appendix B Hyperparameter Optimization

Hyperparameter optimization was performed using Bayesian sweeps (Weights & Biases). Key optimized parameters included:

- Loss ratios (BCE, Dice, Focal)
- Mixup alpha (optimized: 0.3604)
- Augmentation strength (optimized: 1.137)

Sweep configurations and resulting optimal hyperparameters are documented and available in the project’s WandB repository.

References

1. Yazdi, S., Baqersad, J.: Mechanical modeling and characterization of human skin: a review. *J. Biomech.* **130**, 110864 (2022). <https://doi.org/10.1016/j.jbiomech.2021.110864>
2. Foltynski, P., Ladyzynski, P.: Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybernetics Biomed. Eng.* **43**(1), 7–29 (2023). <https://doi.org/10.1016/j.bbe.2022.11.004>
3. Li, F., Wang, C., Liu, X., Peng, Y., Jin, S.: A composite model of wound segmentation based on traditional methods and deep neural networks. *Comput Math Methods Med.* (2018). <https://doi.org/10.1155/2018/4149103>
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation (2015). <https://arxiv.org/abs/1505.04597>
5. Niri, R., Zahia, S., Stefanelli, A.E.A.: Wound segmentation with u-net using a dual attention mechanism and transfer learning. *J. Digit. Imaging* (2025). <https://doi.org/10.1007/s10278-025-01386-w>
6. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers (2021)
7. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Stoyanov D et al (eds) *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, Cham, pp 3–11
8. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818 (2018). https://doi.org/10.1007/978-3-030-01234-2_49
9. Cui, L., Li, D., Yang, X., Liu, C.: Towards reliable healthcare imaging: conditional contrastive generative adversarial network for handling class imbalance in mr images. *J. Biomed. Inf.* (2023)
10. Zimmermann, N., Sieberth, T., Dobay, A.: Automated wound segmentation and classification of seven common injuries in forensic medicine. *Forensic Sci. Int.* (2023). <https://doi.org/10.1007/s12024-023-00668-5>
11. Wang, C., et al.: Fuseg: the foot ulcer segmentation challenge. *Information* **15**(3) (2024). <https://doi.org/10.3390/info15030140>



Wound Area Detection with nnU-Net v2 Architecture

Franciszek Kubala , Piotr Matusiewicz^(✉) , Dominik Mika ,
Dominik Zawlocki , and Zbislav Tabor

AGH University of Krakow, 30-059 Karkow, Poland
pmatusiewicz@student.agh.edu.pl

Abstract. Diabetic foot ulcers are a dangerous complication of diabetes and, if uncared for, can lead to significant health risks. In treatment, wound annotation and measurements are crucial. The development of AI models has the potential to significantly transform medical diagnostic techniques and greatly benefit the daily work of medical staff. With the support of deep learning, we aim to expand on new ways to implement convolutional neural networks (CNNs) in the daily work of medical staff. In this study we used the well renowned nnU-Net 2.0 architecture. The models are trained using custom loss function which improves segmentation accuracy compared to the state of the art approaches. Segmentation was performed using an ensemble of models. For the dataset, we obtained 451 wound pictures with ground truth segmentations of wounds and scale markers and split the dataset into a holdout test part of 15% images and 85% of training/validation images. Various models achieved accuracy of 90.91%, 90.88% and 91.65%. Model ended up with big accuracy and could be used in a medical setting.

Keywords: Wound area segmentation · Convolutional neural networks · Deep learning · nnU-Net

1 Introduction

Diabetes is a chronic metabolic disorder characterized by high blood glucose levels. According to the World Health Organization (WHO) in 2022, 810 million people worldwide were suffering from diabetes, more than 10% of the global population. This number is only going to grow rapidly [1]. There are a lot of possible severe complications of diabetes, but in this study we will specifically focus on diabetic foot ulcers. Diabetic foot ulcers (DFU) are a major complication of prolonged and neglected diabetes. It manifests as an open wound or sore that develops on the foot due to nerve damage (neuropathy) and poor blood circulation (ischemia). Studies show that approximately 15% of diabetes patients develop DFU and 14-24% of them will require amputation of the affected foot due to bone infection or other ulcer-related complications [2]. With that many victims, more sophisticated and accessible methods of segmentation are in much need for doctors and medical staff.

Accurate and precise wound assessment plays a crucial role in the management of DFUs and is typically performed by medical professionals. However this is very time

consuming and inefficient due to the increasing number of patients and shortage of hospital personnel. Ideal solution would be for computer programs to assist with those tasks, and certainly it is becoming more popular to create such assistants. In recent years there has been a rise of AI-based segmentation software for medical purposes. In particular, convolutional neural networks (CNNs) and deep learning algorithms have gained recognition in the biomedical field due to their adaptability and versatility for different types of usages [3]. In recent years there was a rise in interest in wound segmentation. Many of them focused on DFU wounds and various methods were tried [7–9], which differ in measurement accuracy, technology and usage form. For this study we chose specifically nn-Unet v2.0 architecture because of its well-known reputability, scientific achievements and state-of-the-art performance in image segmentation. It allows for great variability and easy handling. It is different from the usual deep learning algorithm as it configures itself based on the dataset and unique structure. Although nnU-Net is designed to work out-of-the-box, it freely allows engineers to modify and enhance its functionality for specific purposes. In our study, we leveraged custom loss functions to further refine segmentation accuracy.

2 Material and Methods

2.1 Dataset

For the purpose of this study, we used 451 wound three-channel optical images with ground truth segmentation masks of wounds and scale markers. We split the dataset into separate training and testing cases. Due to the built-in nnU-net five-fold cross-validation training strategy there was no need for a separate validation dataset since our training dataset was internally split by nnU-net into training and validation subsets (separate split for each of the five folds). The same split of the original dataset into 85% training and 15% hold-out testing cases was used in all experiments. The training and test images datasets consisted of, respectively, 384 and 67 cases. After splitting the dataset into training and testing subsets, these subsets were preprocessed using a dedicated nnU-net preprocessing script, as described in detail in nnU-net documentation.

2.2 The Use of CNN with U-Net Architecture for Image Segmentation

To address the problem of wound area segmentation, we decided to use U-Net architecture. Our approach is based on nnU-Net v2 licensed under the Apache License 2.0. The nnU-Net v2 is the successor of the awarded U-Net CNN that won the International Symposium on Biomedical Imaging in 2015 [3]. U-Net uses an encoder that is connected to the decoder via a bottleneck. Each encoder level performs convolution and down sampling of the image. Then the bottleneck connects the encoder with the decoder. The decoder performs the reverse process, up sampling output of the bottleneck to the original dimension using additional encoder features injected through skip connections. Top level layer extracts features related to the segmentation classes. In the case of the dataset analyzed in the present study the encoder and the decoder are made of four levels.

As a part of the loss function we used Dice coefficients given by following formula:

$$DSC_W = \frac{2 \cdot TP_W}{2 \cdot TP_W + FP_W + FN_W}$$

$$DSC_S = \frac{2 \cdot TP_S}{2 \cdot TP_S + FP_S + FN_S}$$

The W subscript refers to wounds and S subscript refers to scale markers. Wounds and scale markings are represented in the nnU-net framework by different segmentation classes. The TP, FP, TN, FN are explained in Table 1. By default, the nnU-net loss function is a sum of a cross-entropy and a class-averaged Dice coefficient.

Table 1. Table of dice formula variables

Variable	Is classified correctly	Is classified as wound or scale
TP	Yes	Yes
FP	No	Yes
TN	Yes	No
FN	No	No

2.3 Alternative Approaches to Segmenting Wounds and Markers

Three different approaches were tested to solve the problem of wounds and markers segmentation. The first, reference approach, was based on a plain nnU-net, used for training a two-class segmentation model.

The second approach involved training two separate CNNs models: one for wounds and the second for markers. The third approach is focused on training one CNN multi-class segmentation model with a modified loss function. All the trained models were subsequently used for accuracy assessment on the fixed hold-out test set.

In the second approach, the two separate binary segmentation models (one for wounds and the other for scale markers) were applied in sequence to an input image and then, the segmentation outputs of these models were combined to predict the final segmentation of wound and scale markers.

In the third approach a single model was trained using a custom loss function. In particular, we replaced a Dice-coefficient related part of the standard nnU-net loss function by a weighted class-average of F1 coefficient, given by the following formula:

$$F1 = \frac{3}{4} \cdot DSC_W + \frac{1}{4} \cdot DSC_S$$

By that we focused model training on more complex wounds class segmentation.

2.4 Evaluation

We use F1 score to evaluate segmentation results, in particular to assess the quality of automated segmentations compared to ground truth segmentations. The F1 score is given by the following formula:

$$F1 = \frac{3}{4} \cdot F1_W + \frac{1}{4} \cdot F1_S$$

Then F1W and F1S for each image are calculated by:

$$F1_W = \frac{TP_W}{TP_W + \frac{1}{2}(FP_W + FN_W)}$$

$$F1_S = \frac{TP_S}{TP_S + \frac{1}{2}(FP_S + FN_S)}$$

A final score for a test dataset is a mean over scores $F1_i$ calculated for test cases:

$$F1 = \frac{1}{n} \sum_{i=1}^n F1_i$$

We also calculated mean of $F1_W$ and $F1_S$:

$$F1_W = \frac{1}{n} \sum_{i=1}^n F1_{W_i}$$

$$F1_S = \frac{1}{n} \sum_{i=1}^n F1_{S_i}$$

Finally, we calculated standard deviation, minimal, and maximal values of F1, F1W and F1S to get a more comprehensive description of models' performance. The worst cases were also inspected, as described in the results section.

3 Results

For the reference model we got an average F1 score equal to 90,91%, F1W equal to 88,46%, and F1S equal to 98,27%. For the second approach, based on training two separate binary segmentation models, after combining the results of the two models, we got an F1 score equal to 90,88%, F1w to 88,44% and F1s to 98,23%.

For the third approach (modified loss function) we got an F1 score equal to 91,64%, Fw to 89,41% and Fs to 98,32%. Further details are presented in Table 2.

Wounds score of reference model for validation dataset (Figs. 1, 2 and 3):

Figure 4 shows photo of best segmentation result based on F1 score, ground true mask, predicted mask respectively

Figure 5 shows photo of worst segmentation result based on F1 score, ground true mask, predicted mask respectively

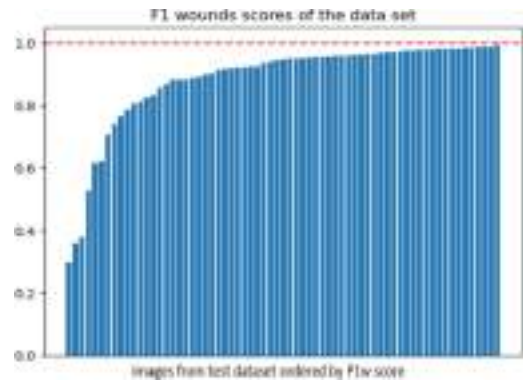


Fig. 1. F1 scores value for each image from the test set of the reference model.

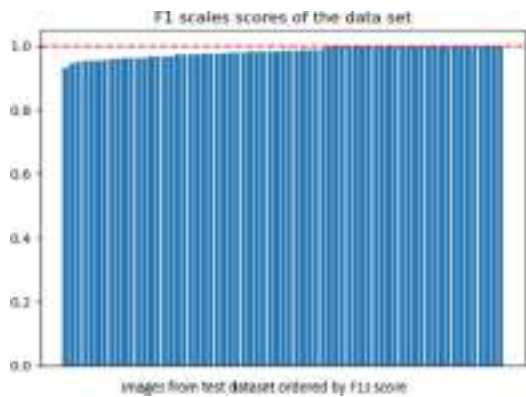


Fig. 2. F1w scores value for each image from the test set of the reference model.

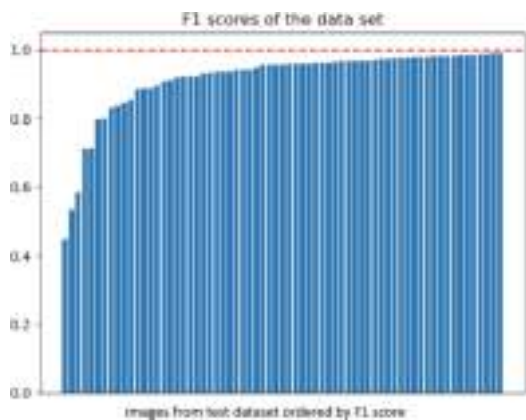
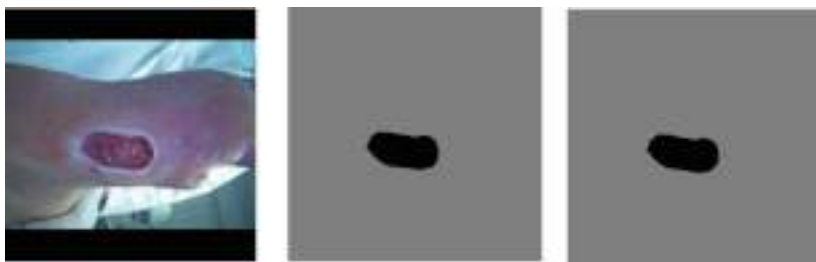


Fig. 3. F1s scores value for each image from the test set of the reference model.

**Fig. 4.****Fig. 5.****Table 2.** Table of model scores

Model	Reference model	Two binary-segmentation models	Model with modified loss function
avg F1	90,91 %	90,88%	91,65%
std F1	11,03%	11,23%	10,51%
max F1	99,08%	98,97%	99,14%
min F1	0,42%	47,23%	44,68%
avg F1 _w	88,46%	88,43%	89,42%
std F1 _w	14,37%	15,03%	14,06%
max F1 _w	99,28%	99,29%	99,29%
min F1 _w	0,22%	29,64%	26,24%
avg F1 _s	98,27%	98,23%	98,33%
std F1 _s	1,80%	1,77%	1,71%
max F1 _s	100%	100,00%	100%
min F1 _s	0,93%	93,15%	93,92%

4 Model Access and Deployment

4.1 Different Approaches

There are different ways to use trained neural networks. During development of our solution, we analyzed its performance for usability purposes. By far the most efficient approach for running automated segmentation using nn-Unet framework is to make predictions on dedicated hardware with a graphics processing unit (GPU), accessing computation infrastructure via an API in a web browser or application. However, this involves a network connection, which could result in significant latency or data loss if the internet connection issues. In addition, time would need to be spent on proper connection and data protection, which issues are outside of the scope of the present paper.

Another approach we considered during our work was to create a self-contained executable file that could be run from a terminal command line. However, compiling a large number of files and data from a trained network into a single executable file makes it extremely difficult to make changes (e.g. update model trainable parameters), as it requires rebuilding the entire application.

4.2 Accessing Trained nn-Unet Model

We ended up creating a lightweight executable file that runs the included Python script customized only to run on the chosen processor and operating system architecture. However, such a solution generates the need to compile the file on the chosen platform - fortunately, thanks to the lightweight nature of this application, we can perform a quick compilation. The advantage of such a solution is that it is comparatively easier to adapt the application to individual needs, and also easy to replace the network model if it is trained on new data.

5 Discussion

In the present paper we demonstrated that wounds can be segmented with satisfactory precision from optical images, using nnU-net framework. Subsequently, measurements, like wound area, diameters, perimeter can be made, using scale markers segmentation. These measurements are crucial for clinical application. While measurements were not a part of the present study, follow up research can include them, in particular to compare manual and automated approaches to wound assessment. This clinically oriented follow up study is necessary to judge whether the precision of segmentation, as achieved in the present study, is sufficient for clinical applications, for example for assessing progress of disease or effectiveness of the applied therapy.

However, even given the results of the present study we may conclude that training a CCN model with the limited dataset in hand resulted in a set of drawbacks. For example, we observed that wounds neighboring ulcers have a highly negative effect on segmentation accuracy. Having a small sample size of some specific types of ulcers resulted in the same negative result. Clearly, such rare cases are not sufficient for training a universal segmentation model.

Given the limitations of the present study following future research should, likely, focus on the following issues:

- To mitigate the problems related to rare cases, hard examples mining techniques can be tried while training the models,
- If possible, the dataset should be extended by the rare cases, keeping in mind that the distribution of cases in the training set must reflect the distribution of cases in clinical settings,
- Instead of training a single, universal segmentation model, the dataset of optical images may be, likely, split into different classes, each class characterized by some distinctive clinical features; then, separate segmentation models should be trained for each class, and - at inference time - a classifier should be first applied to the analyzed case, followed by segmentation with a model selected by the classifier,
- A straightforward extension of the present study is segmentation uncertainty assessment; indeed, given that the training is run using five-fold cross-validation, the five trained models can be used to estimate segmentation uncertainty,
- Finally, pixels are assigned to classes (background, wounds, scale markers) based on maximum likelihood which need not to be the optimal solution for the case in hand; selection of optimal decision thresholds, based e.g. on ROC analysis should be tested.

6 Conclusion

Findings of this study highlight the importance of nnU-Net v2 architecture as very versatile and flexible. During our work it became apparent from the reference model, that even with a small training dataset, making CNN model with base nn-Unet architecture results in high accuracy. On top of that, experimenting and optimizing structure we can achieve even greater accuracy. Training CNN models on every type of wound/irregularities of interest seems viable and productive provided a well selected dataset with masks provided by professionals along with a small group of engineers experienced in nn-Unet architecture.

Acknowledgements. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018094

References

1. Raja JM, Maturana MA, Kayali S, Khouzam A, Efeovbokhan N (2023) Diabetic foot ulcer: a comprehensive review of pathophysiology and management modalities. *World J Clin Cases* 11(8):1684–1693. <https://doi.org/10.12998/wjcc.v11.i8.1684>. PMID:36970004;PMCID:PMC10037283
2. WWW: Web page of the World Health Organization (WHO) about Diabetes. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
3. Devi, M., et al.: A survey of soft computing approaches in biomedical imaging. *J. Healthcare Eng.* 2021.1 (2021): 1563844
4. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation

5. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18(2):203–211
6. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43:17-29. <https://doi.org/10.1016/j.bbe.2022.11.004>
7. Huang HN et al (2022) Image segmentation using transfer learning and Fast R-CNN for diabetic foot wound treatments. *Front Public Health*. 20(10):969846. <https://doi.org/10.3389/fpubh.2022.969846>. PMID:36203688;PMCID:PMC9530356
8. Wang C et al (2024) FUSeg: the foot ulcer segmentation challenge. *Information* 15:140
9. Thotad PN, Bharamagoudar GR, Anami BS (2023) Diabetic foot ulcer detection using deep learning approaches. *Sensors International* 4 (2023), 100210, ISSN 2666-3511. <https://doi.org/10.1016/j.sintl.2022.100210>



Segmentation Challenge: Wounds and Markers in Focus

Piotr Potrzebowski^(✉), Arkadiusz Karbowski, Kamil Sterniuk, Aleksandra Krajna, Zuzanna Cemka, Krystian Opała, and Jacek Rumiński

Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, G. Narutowicza 11/12, 80-233 Gdańsk, Poland
s185416@student.pg.edu.pl

Abstract. Accurate segmentation of wound images is critical for effective diagnosis, treatment planning, and monitoring of wound healing. In this study, we introduce an advanced deep learning framework for the simultaneous segmentation of wounds and scale markers. Our approach leverages state-of-the-art architectures—including U-Net++, DeepLabV3+, and Segformer—combined with pre-trained encoders and diverse transfer learning strategies (full fine-tuning, decoder-only tuning, and partial tuning) to address the challenges posed by limited and heterogeneous clinical datasets. Evaluations conducted on the NBC2025 Challenge dataset demonstrate the effectiveness of our method, with models achieving high F1-scores, including 0.8934 on the test dataset, and competitive mean Intersection over Union (mIoU) values on the validation set. Extensive data augmentation further enhances model generalization, ensuring robust segmentation performance. The promising results underscore the potential of deep learning-based segmentation as a reliable tool for objective wound assessment, thereby supporting improved clinical decision-making and patient care.

Keywords: wound segmentation · deep learning · medical image analysis · convolutional neural networks · transfer learning · semantic segmentation

1 Introduction

A wound is broadly defined as any disruption in the normal structure or function of the skin and underlying tissues. In medical terms, this refers to damage that compromises the cellular, anatomical, or functional integrity of living tissue [1]. Wounds are commonly categorized by their healing trajectory into acute (such as surgical incisions or accidental lacerations) and chronic types (those that fail to heal within the usual period).

Acute wounds are extremely common – one study estimates about 11 million acute wounds occur each year in the United States, with roughly 300,000 cases requiring hospitalization [2]. Chronic wounds, while less frequent, impose a heavy

burden: they are estimated to affect about 1–2% of the population in developed countries at any given time [3].

Traditional wound assessment in clinical practice is largely manual, which is labor-intensive – clinicians measure wound dimensions and estimate tissue types by visual inspection. Automatic wound segmentation using deep neural networks addresses these challenges by providing objective and reproducible analyses of wound images. The deep learning model can learn to outline a wound and identify its component tissues with a high degree of consistency, doing so in seconds and speeding up the assessment process. It also improves accuracy and consistency, as the algorithm applies the same criteria to each image, eliminating differences between observers and reducing human error in measurements.

AI-driven wound assessment enhances clinical outcomes by providing timely, accurate data for optimized treatment, reducing complications. Studies show it improves wound care efficiency, enables personalized treatment, accelerates healing, and lowers costs.

In computer vision, semantic segmentation refers to the process of classifying each pixel in an image into a category, thereby partitioning the image into meaningful regions corresponding to different object classes [4]. Semantic segmentation provides a pixel-accurate understanding of the scene. Modern approaches to semantic segmentation rely heavily on deep learning: for example, Fully Convolutional Networks (FCNs) introduced by Long et al. were pioneering CNN models that enabled end-to-end segmentation of images of arbitrary size [5].

In the medical imaging field, semantic segmentation is especially critical because clinical decisions often depend on precise localization of anatomical structures or lesions. Medical image segmentation is considered one of the most important tasks in medical image analysis. When it comes to wound image analysis, semantic segmentation techniques play a key role in automating wound assessment. By applying semantic segmentation to wound photos, clinicians obtain objective data like wound size, shape, and tissue composition without relying solely on visual estimation [6].

Over the past decade, standard CNN architectures, including the Fully Convolutional Network (FCN) introduced by Long et al. [7] and the widely recognized U-Net proposed by Ronneberger et al. [8], have become foundational for automatic medical image segmentation. The FCN advanced pixel-wise labeling by repurposing convolutional layers for dense prediction [9], while U-Net employed an encoder-decoder design with skip connections to achieve state-of-the-art performance across diverse biomedical tasks [10]. These models have likewise proven effective in wound image analysis, offering consistent and accurate delineation of wound boundaries under varying conditions of lighting and tissue appearance. Although newer architectures that integrate attention mechanisms like transformer- and attention-based modules continue to emerge, these traditional CNN-based methods remain widely utilized baselines and are essential benchmarks for evaluating more advanced segmentation techniques.

Our contributions include, but are not limited to:

- A comprehensive exploration of diverse model architectures and backbones for enhanced wound segmentation.
- The development of robust data augmentation techniques that significantly improve segmentation outcomes.
- A systematic investigation into various transfer learning strategies and their impact on overall segmentation performance.

2 Related Work

Numerous researchers have advanced wound segmentation through innovative deep learning approaches, focusing on architectural improvements, attention mechanisms, and efficient training strategies.

Wang et al. [11] proposed a fully automatic wound segmentation framework based on MobileNetV2. Their method balances computational efficiency with accuracy by incorporating connected component labeling (CCL) for post-processing, achieving a Dice score of 90.47% on a foot ulcer dataset.

Scebba et al. [12] introduced the “Detect-and-Segment” (DS) approach, which unifies wound detection and segmentation. By first localizing the wound and then segmenting the isolated region, the DS method significantly improved performance. For diabetic foot ulcers, DS increased the Matthews’ correlation coefficient (MCC) from 0.29 to 0.85 for U-Net and from 0.35 to 0.82 for ConvNet, with corresponding Dice scores rising from 0.10 to 0.75 and 0.15 to 0.70. Furthermore, evaluation on four independent datasets (covering diverse wound types and body locations) demonstrated a robust improvement in out-of-distribution performance (mean MCC from 0.17 to 0.85), and training was possible with up to 90% less data—with U-Net maintaining performance within a 6.4% MCC range (0.84-0.78) when reducing the training set from 90% to 10%.

Niri et al. [13] proposed a dual attention U-Net model that integrates attention gates (AG) and squeeze-and-excitation (SE) blocks using a VGG16 backbone. This architecture leverages spatial attention to highlight relevant wound regions and channel-wise SE blocks to recalibrate feature maps, enabling multi-level feature refinement. Initially trained on the DFUC dataset (2,000 diabetic foot ulcer images), the model was also evaluated on the WTS (2,230 images), STANDUP (569 images), and FuSeg (1,210 images) datasets. For comparison with other state-of-the-art methods, the WTS dataset achieved a Dice coefficient of 94.10 and an IoU of 89.30.

Addressing the heterogeneity in wound appearances, Oota et al. [14] introduced WSNET, a framework that combines three innovations: a global-local architecture, wound-domain adaptive pretraining (WDAP), and specialized data augmentation. The global-local architecture processes both the entire image ($192 \times 192 \times 3$) and 16 non-overlapping patches ($48 \times 48 \times 3$) in parallel with shared weights, capturing macro-level context and fine-grained details. WDAP on 125,711 unlabeled wound images significantly enhances feature representation compared to ImageNet pretraining. Their experiments with U-Net, LinkNet, PSPNet, and FPN using 17 different backbones showed that the best model

achieved an IoU of 0.713 and Dice of 0.847 on the WOUNDSEG dataset (2,686 images across eight wound types), with Dice scores reaching 0.927 on the AZH Woundcare dataset and 0.956 on the Medetec dataset.

Ramachandram et al. [15] focused on automated wound tissue segmentation using deep neural networks to reduce subjectivity in traditional assessments. Using the Swift Wound Data Set (58 chronic wound images captured under varying lighting conditions and angles), they developed two models: one for overall wound area segmentation (achieving a mean Intersection over Union (mIOU) of 0.8644) and one for tissue segmentation (mIOU of 0.7192). Clinician consensus rated 91% of the tissue segmentation results as good or acceptable, despite low to moderate inter-clinician agreement.

Collectively, these studies illustrate a progression from lightweight, efficient architectures and unified detection-segmentation methods to attention-enhanced and dual-stream frameworks, underscoring the potential of these approaches to improve segmentation performance across diverse wound datasets.

3 Materials and Methods

3.1 Dataset

In the process of training and evaluating our wound segmentation model, we used two different datasets:

3.1.1 NBC2025.Challenge Dataset The first dataset used in the study was provided as part of the IFMBE Scientific Challenge Competition at the NBC 2025 & PCBBE 2025 conference courtesy of the IFMBE Digital Health Division. This dataset comprises wound images and linear scale markers, originally collected within the AutoPlanimator project and described by Foltynski and Ladyżyński (2023) [16].

The dataset consists of two subsets:

- **Training set** containing 371 JPG images with a resolution of 512×512 pixels, accompanied by corresponding reference masks in PNG format,
- **Validation set** comprising 80 JPG images of the same resolution with corresponding reference masks.

The images and their respective masks are organized in a structured directory format, where each image is associated with a reference mask consisting of three classes: black ($RGB = 0, 0, 0$) representing the wound area, gray ($RGB = 127, 127, 127$) for the background, and white ($RGB = 255, 255, 255$) for scale markers.

Each image contains at least one wound. Some images may depict multiple wounds and multiple scale markers, necessitating a segmentation model capable of appropriate generalization. While some images lack a scale marker, every file in the test set contains at least one scale marker and one wound. This dataset serves as a unique resource for evaluating wound and scale marker segmentation algorithms, which is crucial for advancing automated wound image analysis in clinical practice. Figure 1 shows sample images from the described database.

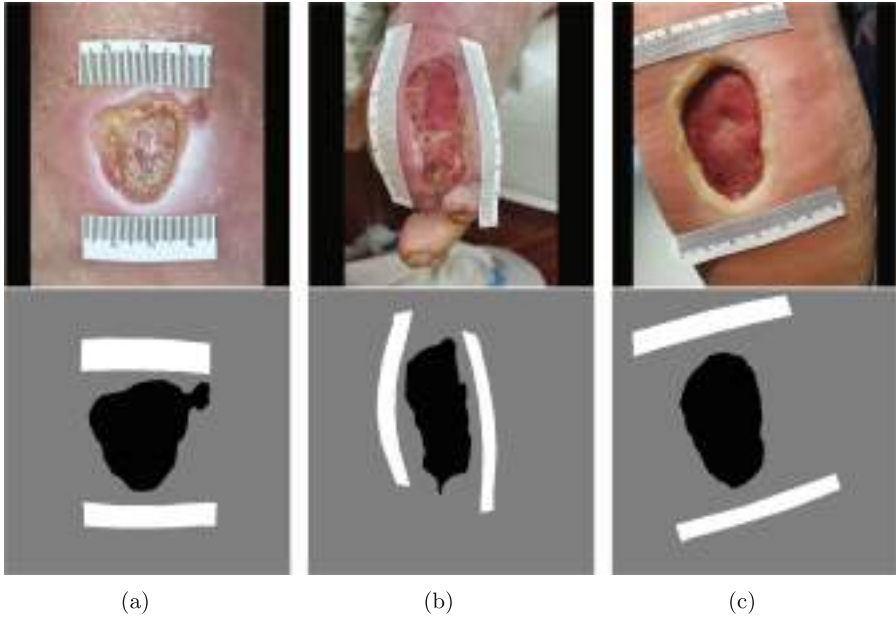


Fig. 1. Images (a), (b), (c) show examples of three wounds from the NBC2025 Challenge dataset. The top row shows the actual wound images, and the bottom row shows the segmentation masks.

3.1.2 Wound Segmentation Images Dataset The second dataset utilized in this study is the Wound Segmentation Images Dataset [17], which compiles wound images from multiple sources including the Medetec Wound Database, the Foot Ulcer Segmentation Challenge (FUSeg) dataset, and the WSNet study. This diverse collection contains wound images captured in real-world clinical settings and is accompanied by refined, manually annotated segmentation masks.

The dataset is divided into two subsets:

- **Training set (2208 images):** Includes images with corresponding reference masks for training segmentation models,
- **Test set (552 images):** Consists of images without reference masks, enabling independent evaluation of segmentation performance.

All images in the dataset have been resized to a resolution of 512×512 pixels, regardless of their original format. For instance, images from FUSeg were already standardized to this resolution, while images from WSNet, which initially had approximately 384×384 pixels, were also adjusted to this size.

Captured during actual patient visits, the images exhibit variability in lighting, perspective, and wound severity.

Additionally, artificial linear markers have been overlaid on the images to tailor the dataset to the segmentation task (Fig. 2).

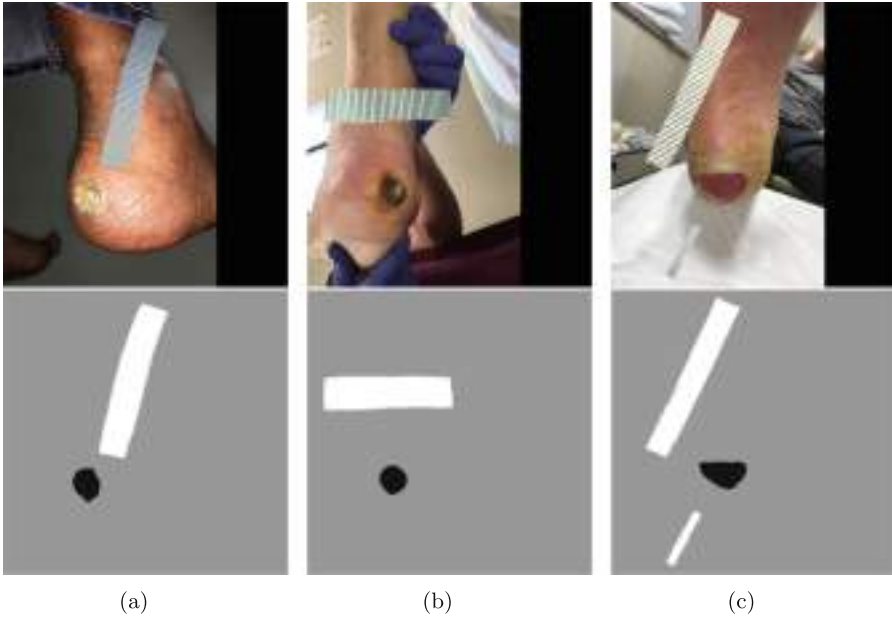


Fig. 2. Images (a), (b), (c) show examples of three wounds from the Wound Segmentation Images Dataset. The top row shows the actual wound images, and the bottom row shows the segmentation masks. Artificial linear markers, specifically added for the segmentation task, are clearly visible in these examples.

3.2 Pre-processing

Deep neural networks require a large amount of training data to achieve optimal performance and prevent overfitting. However, obtaining a sufficient number of training samples is often challenging. To construct a training dataset, images must first be acquired and then annotated. In the case of semantic segmentation tasks, each pixel of the input image must be assigned a correct class label. This process is labor-intensive, and in some cases, data annotation can be highly costly. For instance, accurately labeling medical images requires expensive domain experts, and even acquiring training images can be difficult.

Therefore, appropriate preprocessing procedures were necessary before initiating the training process. Given the relatively small dataset size, data augmentation techniques were employed to enhance model performance and increase dataset diversity.

Image augmentation is the process of generating new training examples from existing ones by applying slight modifications to the original images. For example, an image can be slightly darkened, a region of the original image can be cropped, a new image can be created by mirroring the original, or it can be scaled. In our implementation, we focused on transformations such as:

- **Spatial transforms:**
 - Horizontal and vertical flipping (probability = 50%)
 - Random 90-degree rotations (probability = 50%)
 - Affine transformations, including small translations (up to 6.25% shift), scaling (90%–110%), and rotations (-45° to 45°)
 - Elastic deformation of the image (probability = 10%)
 - Grid distortion with five control points and a distortion limit of 30% (probability = 10%)
- **Color transformations:**
 - Random changes in brightness and contrast (probability = 20%)
 - Hue, saturation, and value adjustments (hue shift limit = 20, saturation shift limit = 30, value shift limit = 20, probability = 20%)
 - Contrast Limited Adaptive Histogram Equalization (CLAHE) for improved contrast (probability = 10%)
- **Noise injection:**
 - Gaussian noise with a standard deviation range of 0.1 to 0.3 (probability = 10%)
 - ISO noise simulation with color shift (0.01–0.05) and intensity variation (0.1–0.5) (probability = 10%)
- **Image quality adjustments:**
 - Gaussian blur (probability = 10%)

These transformations were selected to introduce sufficient variation in the dataset while preserving the essential characteristics of the wound and marker segmentation task. The augmentation pipeline ensures improved model generalization by preventing overfitting to specific imaging conditions.

3.3 Models and Training Strategies

The first strategy, which was extensively utilized, involved using a pre-trained encoder model as a backbone for a segmentation model.

The second approach examined involved pretraining the model weights on a larger but similar dataset before applying transfer learning to the target dataset. In both approaches, the backbone—i.e., the pre-trained encoder—played a crucial role in providing robust feature representations from the start. Three different transfer learning methods were then employed:

1. **Decoder-only approach** – all encoder weights were frozen during model tuning.
2. **Full fine-tuning** – none of the model’s weights were frozen.
3. **Partial approach** – all encoder layers except the last one were frozen.

All of these methods enabled effective adaptation of the model to the segmentation task and allowed for an evaluation of the impact of full training versus knowledge transfer on the final results.

3.4 Evaluation Metrics

3.4.1 F1-Score The evaluation of segmentation models will be conducted based on the F1-score metric. The F1-score is a crucial performance indicator in machine learning, providing a balanced measure of a model's precision and recall.

It represents the harmonic mean of precision and recall, making it particularly useful in segmentation tasks, where it is essential to minimize both false positives (FP) and false negatives (FN).

For each test image, the predicted segmentation mask will be compared with the corresponding ground truth mask. The evaluation process will involve calculating the F1-score separately for the wound region ($F1_w$) and the scale marker region ($F1_s$), followed by computing the overall segmentation score for the image. This is an evaluation method that was specifically defined in connection with the IFMBE Scientific Challenge.

Mathematically, the F1-score is defined as [19]:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- Precision = $\frac{TP}{TP+FP}$ (the proportion of correctly identified wound pixels among all detected wound pixels),
- Recall = $\frac{TP}{TP+FN}$ (the proportion of actual wound pixels correctly detected by the model).

For segmentation evaluation, the wound segmentation score ($F1_w$) and the scale marker segmentation score ($F1_s$) are computed as follows:

$$F1_w = \frac{2 \times TP_w}{2 \times TP_w + FP_w + FN_w}$$

$$F1_s = \frac{2 \times TP_s}{2 \times TP_s + FP_s + FN_s}$$

The final F1-score for each test image ($F1_i$) is the average of these two scores:

$$F1_i = \frac{3F1_w + F1_s}{4}$$

To assess the overall performance of a segmentation model, the mean F1-score will be computed across all n test images:

$$\text{Mean } F1 = \frac{1}{n} \sum_{i=1}^n F1_i$$

This mean F1-score will serve as the primary ranking criterion for evaluating and comparing the segmentation models. The F1-score ranges from 0 to 1, where 1 indicates a perfect model with high precision and recall, while 0 represents the worst possible performance.

3.4.2 Intersection over Union (IoU) Metric The second metric used to assess the quality of segmentation is IoU (Intersection over Union). IoU measures the overlap between the predicted bounding box (or segmentation mask) and the ground truth. It is calculated as the ratio of the intersection area to the union area [19]:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

where: - A is the predicted region, - B is the ground truth region, - $|A \cap B|$ is the area of intersection, - $|A \cup B|$ is the total area covered by both regions.

The value range oscillates between 0 and 1, where 0 means no overlap, and 1 represents complete coverage (perfect match).

4 Results

4.1 Implementation and Hardware Details

The models were trained using Python-based software that utilizes libraries and frameworks such as PyTorch and Segmentation Models PyTorch (SMP) [18]. The experiments were conducted on a DGX Station. All models were trained using the Lovsz loss function and the Stochastic Gradient Descent optimizer with a momentum of 0.9 and a learning rate of 0.001.

The experiments were performed using 11 model architectures, such as UNet, UNet++, and SegFormer, with 16 pretrained encoders, including ResNet152, SE-ResNeXt101 (32×4d), and MIT-b5. From these models, UNet++, SegFormer, DeepLabV3+, and UPerNet were chosen for the transfer learning study along with specifically selected encoders. Each of these models, with the chosen encoders, was trained using all three transfer learning approaches. The full list of utilized models is shown in Table 1.

4.2 Utilizing Pretrained Encoders

The best results obtained using pretrained encoders, along with other approaches, are shown in Table 2, for results evaluated by us on the validation dataset, and in Table 3 evaluated by the Organizing Committee on the test dataset. The authors observed that this approach was generally insufficient for achieving high-quality object segmentation on the given dataset. The only exception to this rule was the UNet++ model, which produced desirable results when using the ResNeXt50 (32×4d), ResNet152, and ResNeXt101 (32×8d) encoders.

To provide a qualitative assessment of segmentation performance, we present visualizations of the results for several top-performing models. These models were selected based on their F1-score performance on the validation dataset.

For each model, we performed inference on a set of representative validation samples and generated side-by-side comparisons between:

1. The original input image (unnormalized for better visibility),

Table 1. Complete list of all models and encoders used in the study.

Category	Models and Encoders
All architectures used	Unet, Unet++, FPN, PSPNet, DeepLabV3, DeepLabV3+, Linknet, MAnet, PAN, UPerNet, Segformer
All encoders used	ResNet34, ResNet101, ResNet152, ResNext50 (32x4d), ResNext101 (32x8d), SE ResNet152, SE ResNext101 (32x4d), Efficientnet-b2, Efficientnet-b5, Efficientnet-b6, Timm-Efficientnet-b4, Timm-Efficientnet-b7, MIT-b0, MIT-b1, MIT-b2, MIT-b5
Architectures used in transfer learning	Unet++, DeepLabV3+, UPerNet, Segformer
Encoders used in transfer learning	ResNet152, ResNext50 (32x4d), ResNext101 (32x8d), SE ResNext101 (32x4d), SE ResNet152, MIT-b5, Timm-Efficientnet-b7

- 2. The predicted segmentation mask generated by the model,
- 3. The corresponding ground truth mask.

The results are presented in Fig. 3, illustrating the segmentation performance of different models on sample images.

4.3 Transfer Learning

Among the three transfer learning methods examined on the dataset, full fine-tuning produced the best results. This technique significantly improved the performance of most of the evaluated models, particularly DeepLabV3+ and SegFormer. The highest results were achieved by combining full fine-tuning with UNet++ and SE ResNext101 (32x4d) or ResNeXt50 (32x4d). However, these improvements were only marginal compared to the results obtained without additional transfer learning for UNet++. A slightly larger difference was observed on the test dataset, with an improvement of more than 0.03 in the F1 score.

The best results obtained during evaluation on the validation and test datasets are presented in Tables 2 and 3, respectively.

The visualizations of the results are presented in Figs. 4 and 5, illustrating the segmentation performance of DeepLabV3 and UNet++ on sample images.

Table 2. Best results obtained using different approaches evaluated on validation dataset. B denotes the encoder-only pretrained approach, D represents the decoder-only approach, F stands for full fine-tuning, and P indicates the partial fine-tuning approach.

Model	Encoder	Method	Score	mIoU
UNet++	SE ResNext101 (32x4d)	F	0.8636	0.8798
UNet++	ResNext50 (32x4d)	F	0.8635	0.8788
UNet++	ResNext50 (32x4d)	B	0.8618	0.8700
DeepLabV3+	ResNext101 (32x8d)	P	0.8597	0.8715
DeepLabV3+	ResNext101 (32x8d)	F	0.8579	0.8687
UNet++	SE ResNext101 (32x4d)	P	0.8576	0.8768
UNet++	SE ResNext101 (32x4d)	D	0.8573	0.8767
DeepLabV3+	ResNext101 (32x8d)	F	0.8571	0.8712
DeepLabV3+	ResNext101 (32x8d)	F	0.8540	0.8678
UNet++	ResNet152	B	0.8507	0.8686
UNet++	ResNext101 (32x8d)	B	0.8507	0.8615
UNet++	SE ResNet152	B	0.8507	0.8615
UNet++	ResNext101 (32x8d)	F	0.8488	0.8607
UNet++	ResNext50 (32x4d)	D	0.8486	0.8707
Segformer	MIT-b5	F	0.8456	0.8543
DeepLabV3+	Timm-Efficientnet-b7	F	0.8162	0.8350

Table 3. Best results obtained using different approaches on test dataset. B denotes the encoder-only pretrained approach, D represents the decoder-only approach, F stands for full fine-tuning, and P indicates the partial fine-tuning approach.

Model	Encoder	Method	Score
UNet++	SE ResNext101 (32x4d)	F	0.8934
UNet++	ResNext50 (32x4d)	F	0.8800
UNet++	ResNext50 (32x4d)	B	0.8606

5 Discussion

The experimental results provide valuable insights into the impact of model architectures, encoder choices, and transfer learning strategies on the segmentation quality of wound images and scale markers. The analysis shows that both the design of the architecture and the strategy used for adapting pre-trained weights significantly affect the final segmentation performance. In terms of architecture and encoder performance, the UNet++ model consistently outperformed other models, particularly when combined with SE ResNext encoders. For instance, the UNet++ model with the SE ResNext101 (32x4d) encoder achieved the high-

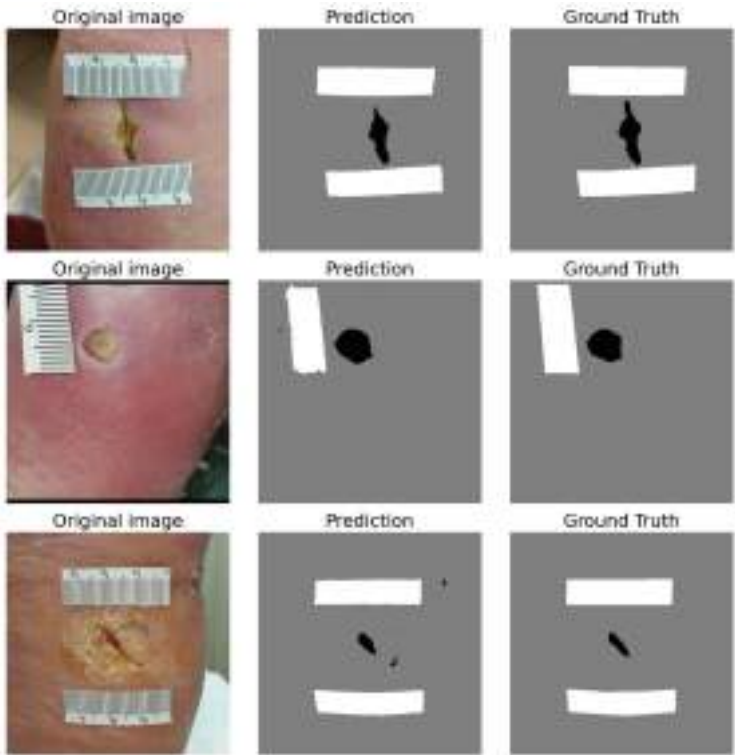


Fig. 3. Visualization of segmentation results for UNet++ with ResNext50 (32x4d) encoder.

est F1-score of 0.8636 and an mIoU of 0.8798 on the validation set, as well as an F1-score of 0.8934 on the test set. This superior performance can be attributed to the model's nested skip connections, which allow for more flexible feature fusion, and the SE ResNext101 encoder's unique combination of a split-transform-merge strategy with channel-wise attention, enabling more effective feature extraction. Similarly, the UNet++ model with the ResNext50 (32x4d) encoder produced nearly identical validation results—an F1-score of 0.8635 and an mIoU of 0.8788—and achieved a strong test F1-score of 0.8800. These outcomes suggest that the cardinality introduced by ResNext architectures is particularly effective for capturing the complex features of wound images, and that the lighter ResNext50 model can achieve comparable results to its deeper counterpart, striking a good balance between model capacity and computational efficiency. In contrast, the DeepLabV3+ architecture with the ResNext101 (32x8d) encoder showed slightly lower performance, with F1-scores ranging from 0.8571 to 0.8597. This indicates that the atrous spatial pyramid pooling in DeepLabV3+ may not provide as much benefit for wound segmentation as the nested skip connections in UNet++, and that the deeper ResNext101 encoder might be prone

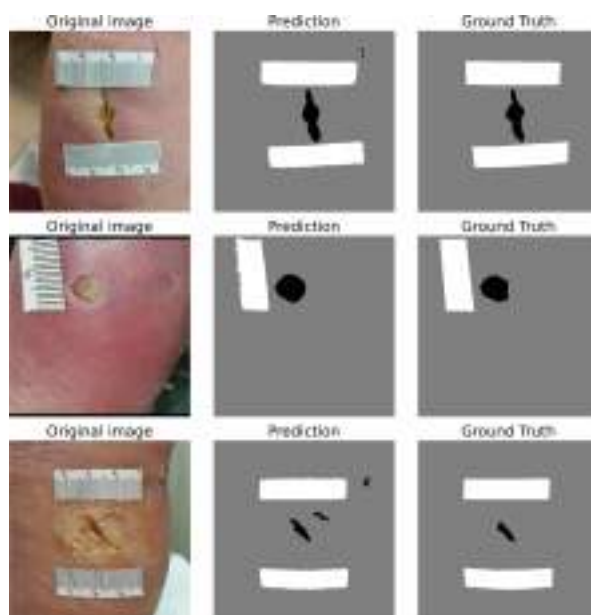


Fig. 4. Visualization of segmentation results for DeepLabV3+ with ResNext101 (32x8d) encoder pretrained on Wound Segmentation Images Dataset.

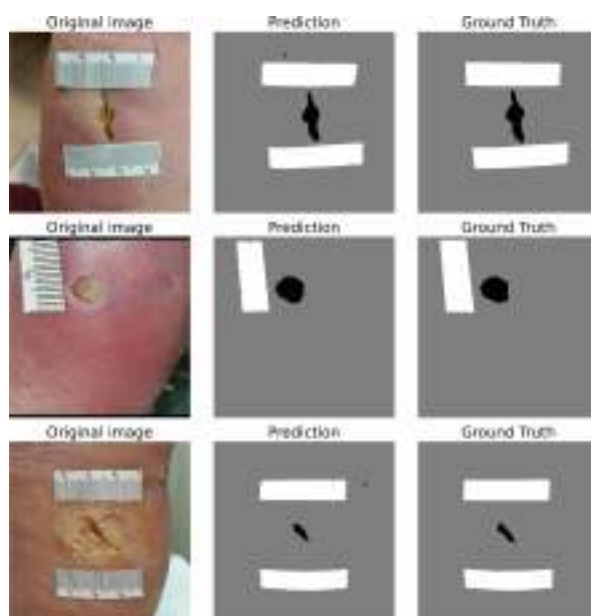


Fig. 5. Visualization of segmentation results for UNet++ with ResNext50 (32x4d) encoder pretrained on Wound Segmentation Images Dataset.

to overfitting on our relatively small dataset. The impact of transfer learning strategies also varied across architectures. Full fine-tuning yielded the best results for most models, particularly improving the performance of DeepLabV3+ and SegFormer. For example, UNet++ with the SE ResNext101 (32x4d) encoder achieved an F1-score of 0.8636 with full fine-tuning, compared to 0.8618 when using only the pre-trained encoder. For DeepLabV3+, partial fine-tuning—where only the last encoder layer and the decoder are updated—slightly outperformed full fine-tuning, with an F1-score of 0.8597 versus 0.8579. This suggests that the pre-trained encoder weights capture general features that remain relevant for wound segmentation, and that fine-tuning only the higher-level layers allows the model to adapt to the task while retaining useful low-level features. Notably, the UNet++ architecture demonstrated high performance even with an encoder-only approach; for instance, UNet++ with the ResNext50 (32x4d) encoder, when relying solely on the pre-trained encoder, achieved an F1-score of 0.8618 on the validation set and 0.8606 on the test set, which is competitive with the fully fine-tuned models.

In comparison with related work, our experimental findings reveal that our approach demonstrates competitive performance. For instance, Wang et al. [11] reported a Dice score of 90.47% using a MobileNetV2-based framework on a foot ulcer dataset, whereas our best-performing model—UNet++ with the ResNext50 (32x4d) encoder in full fine-tuning—achieved an F1-score of 0.8635 and an mIoU of 0.8788. Similarly, Niri et al. [13] and other studies employing attention mechanisms have achieved Dice scores up to 94.10 with IoU values around 89.30. Although our scores are slightly lower than these peak results, it is important to note that direct comparisons are challenging due to differences in datasets, evaluation metrics, and segmentation tasks.

Overall, these findings underscore the importance of carefully selecting both the architecture and transfer learning strategy for wound image and scale marker segmentation. By leveraging advanced architectures like UNet++ combined with robust encoder designs and tailored fine-tuning approaches, highly competitive segmentation performance can be achieved even on challenging datasets.

6 Conclusions

In this study, four deep learning-based approaches to wound segmentation were proposed and analyzed. Experimental results demonstrated that the developed models achieve high performance, with F1-score values ranging from 0.833 to 0.864 when evaluated on the validation dataset and ranging from 0.861 to 0.893 on test dataset. These findings confirm that the applied methods are effective even with a limited dataset, which is particularly significant in clinical applications where access to large, annotated datasets is often restricted.

The developed models provide a solid foundation for further research on automated wound segmentation, supporting the advancement of diagnostic and wound monitoring systems. Their application can significantly facilitate the assessment of wound size and progression, which is crucial for effective patient treatment.

Several directions for future research should be considered. Expanding the dataset by increasing the number of images could improve model generalization, especially under varying lighting conditions and across different wound types. Incorporating attention mechanisms may enhance segmentation precision. Additionally, leveraging larger, multimodal image datasets for pretraining techniques could better adapt models to real-world clinical conditions.

Conducting studies on data obtained directly from medical facilities would enable model validation in practical clinical scenarios, providing valuable insights for developing an optimal solution.

In conclusion, the results of this study indicate that deep neural networks offer a promising solution for wound segmentation. Further refinements could enhance model accuracy and facilitate integration with medical diagnostic tools, potentially improving treatment outcomes for patients with chronic wounds.

Acknowledgement. This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology.

References

1. Nagle SM, Stevens KA, Wilbraham SC (2025) Wound Assessment. [Updated 2023 Jun 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 January. <https://www.ncbi.nlm.nih.gov/books/NBK482198/>
2. Demidova-Rice TN, Hamblin MR, Herman IM (2012) Acute and impaired wound healing: pathophysiology and current methods for drug delivery, part 1: normal and chronic wounds: biology, causes, and approaches to care. *Adv Skin Wound Care*. 25(7):304–14. PMID: 22713781; PMCID: PMC3428147. <https://doi.org/10.1097/01.ASW.0000416006.55218.d0>
3. Falanga V, et al (2022) Chronic wounds. *Nat. Rev. Dis. Primers*. 21; 8(1), 50. PMID: 35864102; PMCID: PMC10352385. <https://doi.org/10.1038/s41572-022-00377-3>
4. Luo H, Bao J, Wu Y, He X, Li T (2023) SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: Proceedings of the international conference on machine learning (ICML)
5. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. arXiv preprint [arXiv:1411.4038](https://arxiv.org/abs/1411.4038). <https://doi.org/10.48550/arXiv.1411.4038>
6. Le DTP, Pham TD (2023) Unveiling the role of artificial intelligence for wound assessment and wound healing prediction. *Explor Med* 4:589–611. <https://doi.org/10.37349/emed.2023.00163>
7. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. arXiv preprint [arXiv:1411.4038](https://arxiv.org/abs/1411.4038). <https://doi.org/10.48550/arXiv.1411.4038>
8. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. arXiv preprint [arXiv:1505.04597](https://arxiv.org/abs/1505.04597). <https://doi.org/10.48550/arXiv.1505.04597>

9. Shelhamer E, Long J, Darrell T (2016) Fully convolutional networks for semantic segmentation. arXiv preprint [arXiv:1605.06211](https://arxiv.org/abs/1605.06211). <https://doi.org/10.48550/arXiv.1605.06211>
10. Alabdulhafith M, et al (2024) Automated wound care by employing a reliable U-Net architecture combined with ResNet feature encoders for monitoring chronic wounds. *Front Med* 11. <https://doi.org/10.3389/fmed.2024.1310137>
11. Wang C, Anisuzzaman DM, Williamson V et al (2020) Fully automatic wound segmentation with deep convolutional neural networks. *Sci Rep* 10:21897. <https://doi.org/10.1038/s41598-020-78799-w>
12. Scebbba G, et al (2022) Detect-and-segment: A deep learning approach to automate wound image segmentation. *Inform Med Unlocked*, vol 29 (2022). 100884. ISSN 2352–9148. <https://doi.org/10.1016/j.imu.2022.100884>
13. Niri R, Zahia S, Stefanelli A et al (2025) Wound segmentation with U-Net using a dual attention mechanism and transfer learning. *J Digit Imaging Inform Med*. <https://doi.org/10.1007/s10278-025-01386-w>
14. Oota SR, Rowtula V, Mohammed S, Liu M, Gupta M (2023) WSNet: towards an effective method for wound image segmentation. In: 2023 IEEE/CVF winter conference on applications of computer vision (WACV), Waikoloa, HI, USA, pp 3233–3242 (2023). <https://doi.org/10.1109/WACV56688.2023.00325>
15. Ramachandram D, Ramirez-GarciaLuna JL, Fraser RDJ, Martínez-Jiménez MA, Arriaga-Caballero JE, Allport J (2022) Fully Automated Wound Tissue Segmentation Using Deep Learning on Mobile Devices: Cohort Study. *JMIR Mhealth Uhealth* 10(4):e36977. PMID: 35451982; PMCID: PMC9077502. DOI: <https://doi.org/10.2196/36977>
16. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43:17–29. <https://doi.org/10.1016/j.bbe.2022.11.004>
17. Leoscode, Wound Segmentation Images Dataset, Kaggle. https://www.kaggle.com/datasets/leoscode/wound-segmentation-images?select=data_woundseg
18. Iakubovskii P (2019) Segmentation Models Pytorch, GitHub repository. <https://github.com/qubvel/segmentationmodels.pytorch>
19. Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging* 15: 1–28. Springer



Automated Segmentation of Wound and Scale Markers Areas Using a U-Net-Based Neural Network

Zuzanna Błaszczowska¹ , Karolina Dzwonkowska¹ , Piotr Stefański¹ ,
and Piotr M. Szczypiński²

¹ Lodz University of Technology, Lodz, Poland
240547@edu.p.lodz.pl

² Institute of Electronics – Affiliated with the Faculty of Electrical, Electronic, Computer and
Control Engineering, Lodz University of Technology, Lodz, Poland
piotr.szczypinski@p.lodz.pl

Abstract. This paper presents a method for automated segmentation of wound areas and scale markers in medical images using a U-Net-based neural network. The model was trained on a dataset of 371 original images, augmented through rotations to increase its generalization ability. Training was conducted over 100 epochs, with dynamic learning rate adjustments to stabilize the process. The model's performance was evaluated using accuracy, loss metrics, and confusion matrices, with segmentation quality assessed through Dice-Sørensen and Jaccard indices. Results show good performance in segmenting wound areas and scale markers, with the model achieving a median Dice-Sørensen coefficient of 0.7068 for wounds and 0.9397 for scale markers. Future work may include expanding the training dataset and exploring the use of classification-oriented neural networks as encoders to further improve segmentation accuracy.

Keywords: Medical image segmentation · U-Net · Deep learning · Wound detection · Scale markers localization · Convolutional neural networks (CNNs) · Dice-Sørensen coefficient · Jaccard index · Data augmentation

1 Introduction

Identifying regions of interest in images is a crucial aspect of visual analysis, particularly in medical diagnostics. The development and application of automated methods for detecting and localizing pathological changes are essential for early diagnosis and the implementation of appropriate treatment. In the case of tissue damage, determining both the etiology and the size of a wound is critical for selecting the most suitable therapy and accelerating the healing process.

Image segmentation is a key stage in image analysis, aimed at dividing an image into smaller, homogeneous regions corresponding to anatomical structures or pathological areas. There is no single, generalized segmentation algorithm—methods are typically selected and adapted based on the specific image class and the problem at hand. The

development of automatic image segmentation tools often relies on machine learning techniques, which identify image features that distinguish the regions of interest from the background. Recently, deep convolutional neural networks have emerged as a powerful approach for adapting segmentation models to specific image classes, yielding highly accurate results.

The evolution of these methods began with neural networks designed for image classification, which categorize images based on their content. An example of such a classification model is its application in assisting with the etiology determination of in vivo burn injuries in medical imaging diagnostics [1]. However, merely classifying images of wounds or burns is insufficient without explicitly identifying the affected areas on the skin's surface. Therefore, another category of neural networks is used for image segmentation, enabling the delineation of specific regions within medical images [2–4]. Since the objective of this study is to develop an image segmentation method for wound analysis, neural network architectures specifically designed for segmentation tasks will be employed.

One of the most effective architectures used in medical image segmentation is U-Net (Fig. 1). This fully convolutional network (FCNN) has a symmetric structure, consisting of an encoder that extracts key features and a decoder that reconstructs the segmented image [5]. Originally designed for biomedical applications, U-Net is particularly well-suited for analyzing wound images and tissue damage [6, 7]. In this context, leveraging this model, along with deep learning techniques, appears to be a highly promising and well-justified approach.

One of the main limitations of convolutional neural networks is their reliance on large training datasets. The substantial number of tunable convolutional filter weights required for feature extraction necessitates training sets comprising thousands or even hundreds of thousands of sample images to achieve proper generalization. This poses a significant challenge in medical image analysis, where the available dataset size may be insufficient. A network trained on a dataset that is too small or non-representative may overfit the training data, resulting in high segmentation or classification accuracy on training images but poor generalization to new cases.

Several approaches exist to mitigate this issue. First, data augmentation techniques can be employed to artificially expand the training dataset by generating images that closely resemble existing ones. Data augmentation, which includes geometric transformations, color space modifications, image mixing, and methods based on generative adversarial networks (GANs), plays a crucial role in increasing the size and diversity of the training set, ultimately improving segmentation performance [8, 9].

A second approach involves reducing the number of trainable network parameters, thereby decreasing model complexity and reducing susceptibility to overfitting. This can be achieved by reducing the number and size of convolutional filters, decreasing the number of network layers, and incorporating dropout layers.

In supervised learning for convolutional networks, it is essential to prepare segmentation masks—label images that indicate the expected segmentation output. These masks must be created with great precision. If a background mask unintentionally includes part of the pathological region, or if a lesion mask partially overlaps with the background,

the network may learn incorrect features. Feature values computed in such misannotated regions will deviate from expected values, introducing outliers. Careful verification of whether the mask correctly aligns with the lesion area is another crucial strategy for improving the performance of neural networks, potentially enhancing segmentation accuracy [10].

Some studies indicate that models based on fully convolutional neural networks (FCNNs), particularly in interactive segmentation approaches, can significantly improve result precision by incorporating user input [11]. Semi-automated segmentation methods allow users to provide additional guidance, such as marking regions of interest, thereby enhancing algorithm effectiveness even in challenging cases. However, in this study, we assume a fully automated approach. Therefore, neural networks requiring user interaction will not be considered.

The objective of this study is to develop an image segmentation method for identifying wound areas and localizing scale markers used as references for estimating their sizes. The training datasets, including masks indicating wound areas and marker locations, were provided as part of the “Segmentation Challenge: Wounds and Markers in Focus.” To address the problem in the challenge, we employed a neural network based on the U-Net architecture. To achieve high segmentation accuracy for wound images, the original U-Net model was adapted. To reduce the network’s susceptibility to overfitting, the number of filters was decreased, and dropout layers were introduced. Additionally, the mask shapes and their alignment with the targeted wound areas were carefully verified within the training dataset. Data augmentation techniques were also applied. All these modifications are described in detail in the Methods section.

The proposed segmentation method for wound images is designed to accurately delineate wound boundaries, facilitating a precise assessment of disease severity. Quantitative and qualitative results that verify the effectiveness of this approach are presented in the Results section, with further discussion and final conclusions provided in the Discussion and Conclusions section.

2 Materials and Methods

2.1 Data Sets Characteristics

The dataset used in this study was provided as part of the IFMBE Scientific Challenge Competition, organized during the 20th Nordic-Baltic Conference on Biomedical Engineering and the 24th Polish Conference on Biocybernetics and Biomedical Engineering (NBC 2025 & PCBBE 2025). The dataset consisted of 451 wound images along with their corresponding segmentation masks (ground truth), divided into two subsets. The training set comprised 371 color images in JPG format with a resolution of 512×512 pixels, each accompanied by a segmentation mask in grayscale and PNG format. The validation set contained 80 images with identical parameters, along with their respective masks.

Each image in both the training and validation sets contains three distinct classes: wound region, background, and scale markers. In the mask images, these regions are represented by specific grayscale values: black (0, 0, 0) for wounds, white (255, 255, 255) for scale markers, and medium gray (127, 127, 127) for the background. The alignment

between segmentation masks and corresponding regions in the color images was visually verified and found to be precise, ensuring high annotation quality.

2.2 U-Net Adaptation

To identify wound areas and scale markers, a U-Net neural network was employed, as its architecture is specifically designed for medical image segmentation. Moreover, this network enables high segmentation accuracy even with small datasets, which is often the case with medical data [12]. The model takes as input an image with specified height, width, and number of channels, and produces an output image of the same dimensions. To tailor the network to the specified problem the output layer was adapted to classify pixels into three categories indicating them with three gray levels: 0 for background, 1 for wound, and 2 for scale markers.

The network consists of two symmetrical pathways arranged in a U-shape and connected through concatenation layers. The image analysis process begins in the contracting path, which extracts features from the image while reducing its dimensionality. The expanding path then reconstructs the image back to its original resolution. The concatenation layers connect the two pathways, enabling precise localization of extracted features and ultimately contributing to the generation of an accurate segmentation map.

In the convolutional layers, the rectified linear unit (ReLU) activation function was applied due to its effectiveness in U-Net architectures [2]. ReLU returns the input value if it is positive; otherwise, it outputs zero, which helps mitigate the vanishing or exploding gradient problem and enhances computational efficiency. He Normal initialization was used for convolutional kernel weights, as it works well with ReLU. This method samples weights from a normal distribution with variance dependent on the number of neurons in the layer, preventing excessively large or small activation values, thus stabilizing the training process [13]. Padding was applied after convolution to maintain the same input and output dimensions. In the contracting path, a max-pooling layer follows the second convolutional layer to reduce input size. Each module of the expanding path begins with a transposed convolution layer, which upsamples data back to the input size.

Several modifications were introduced to the original U-Net model [12] to optimize the learning process. The number of filters was reduced to one-fourth of the original (from 64 to 16 in the first convolutional layer), significantly reducing computation time while maintaining high classification accuracy [3]. This decision was inspired by previous studies where models with fewer filters achieved a Dice-Sørensen index exceeding 97% [14]. In the referenced study, despite reducing the number of filters, segmentation accuracy for liver regions in CT scans remained high.

To prevent overfitting, dropout layers were added between convolutional layers in both pathways. In the first two layers of the contracting path, 10% of weights were dropped, increasing to 20% in the next two layers. In the latent space (bottleneck) region of the network, where both pathways are directly connected, the dropout rate was raised to 30%. In the expanding path, the dropout rate mirrored that of the corresponding layers in the contracting path.

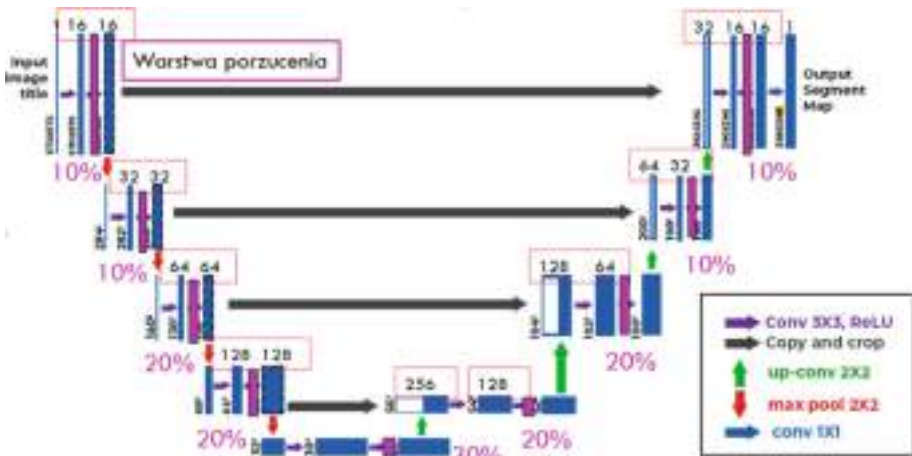


Fig. 1. U-Net architecture with modifications [12].

2.3 Training Process

Initially, the model was trained on 371 original images. However, this number of training examples proved to be too small to achieve high-quality segmentation results on the validation set images. Therefore, the training set was augmented by including rotated copies of the original images, rotated by 90° , 180° , and 270° . As a result, the size of the training dataset increased to over 1200 images. The U-Net training process was then conducted on the entire augmented training dataset. This approach enhanced the model's generalization capability and reduced the risk of overfitting.

The model was trained for 100 epochs using a dynamic learning rate adjustment mechanism, which controlled the rate at which the network's weights were updated during training. Initially, the learning rate was set to 0.001 and was reduced to 0.0001 after 20 epochs. This adjustment helped stabilize the training process and allowed for slower, more detailed optimization in later epochs.

To further mitigate overfitting, the model's performance was continuously monitored on the validation set, and only the best-performing model—determined by the lowest validation loss—was saved. This approach ensured that the model did not overfit to the training set, maintaining its ability to generalize to unseen data.

In addition to the primary model, a secondary model was implemented to test whether the model could learn more effectively using a diverse set of augmented images. This model utilized an augmented dataset consisting of images rotated by 90° , 180° , and 270° degrees. For each of these rotated images, a smaller random crop, approximately one-quarter of the original image size, was extracted from a random starting point. These cropped sections were then resized to a 512×512 format. The model was trained on this augmented dataset, which contained over 3000 images. This approach aimed to enhance the model's learning process and improve its ability to generalize across various transformations.

2.4 Segmentation Quality Metrics

As the training process progresses, changes in accuracy and loss functions are recorded for both the training and validation sets. Accuracy is the number of pixels correctly assigned by the neural network to the background and the region of interest, compared to the reference mask. The binary cross-entropy loss function was chosen to evaluate segmentation quality. It minimizes the difference between the predicted probabilities and the true label values, making it particularly effective for datasets with a high imbalance between the number of pixels in the region of interest and the background [15].

Additionally, the model weights are saved when the lowest loss value is achieved on the validation set. This allows for both tracking and evaluating the entire training process, as well as saving the network state that enables the most accurate segmentation of images.

Using the trained neural network, predictions for wound and scale marker areas were made on the validation set images and compared to the reference masks. Based on the comparison of the segmentation results with the reference masks, confusion matrices were computed. These matrices present four values: true positives (TP), which represent the number of pixels common to both the mask and those identified by the network; false negatives (FN), which are pixels incorrectly omitted by the network; and false positives (FP), which are pixels wrongly identified by the network as belonging to the target area. The number of true negatives (TN)—correctly identified background pixels—is not included in the assessment of segmentation performance, as the size of the background is not relevant.

It is important to note that the network allows for the segmentation of two different types of areas. Therefore, separate confusion matrices were recorded for the prediction of wounds and scale markers.

From the confusion matrix, several classification quality metrics were calculated: True Positive Rate (TPR), Positive Predictive Value (PPV), the Jaccard index, and the Dice-Sørensen coefficient. TPR (1), also known as sensitivity, is the ratio of true positives to all positives.

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

PPV (2), or precision, describes the model's ability to correctly classify positive cases, calculated as the ratio of true positives to the sum of all cases (pixels) classified as positive.

$$PPV = \frac{TP}{TP + FP} \quad (2)$$

The Jaccard (3) and Dice-Sørensen (4) coefficients measure the similarity between the original mask images and the predicted segmentation maps. For both coefficients, a value close to 1 indicates high similarity between the sets, reflecting accurate segmentation results.

$$j = \frac{TP}{TP + FP + FN} \quad (3)$$

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (4)$$

3 Results

Figure 2 presents the segmentation performance of the network throughout the training process across subsequent epochs. Graph (a) shows the accuracy for both the training and validation sets, while graph (b) displays the loss function values. The training was halted after 86 epochs due to a noticeable increase in loss on the validation set, indicating that the model began overfitting and excessively adapting to the training data, thereby losing its generalization ability.

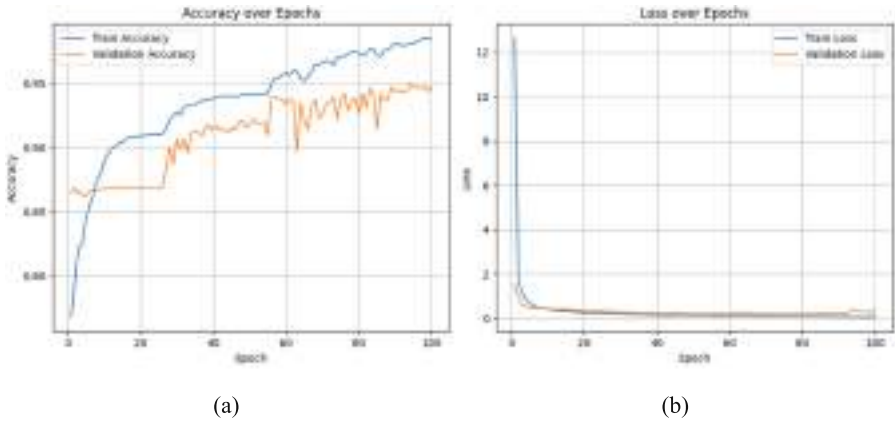


Fig. 2. (a) Plot of accuracy and (b) plot of loss function versus number of epochs for training and validation set.

Figure 3 illustrates the segmentation results for a sample image of wounds and scale markers. The original image (a) is compared with the segmentation results for the wound (b) and the scale markers (c). A comparison with the reference masks (ground truth) allows for a visual assessment of segmentation quality, highlighting correct results (green area) and two types of errors, indicated by red and blue.

To quantitatively evaluate the segmentation results, the Dice-Sørensen and Jaccard metrics were used. These metrics provide a comprehensive assessment of the agreement between the predictions and the actual masks, with values close to 1 indicating high segmentation quality and values near 0 revealing poor performance.

Figure 4 presents a violin plot illustrating the distribution of quality metrics for wound areas. The plots were generated from metrics obtained for 80 images from the validation set. The horizontal axis represents metric values ranging from 0 to 1, while the width of the plot along the vertical axis corresponds to the number of images with a given metric value. Each distribution includes an embedded box plot. The width of the violin plots reflects the density distribution of images for each metric value. The median

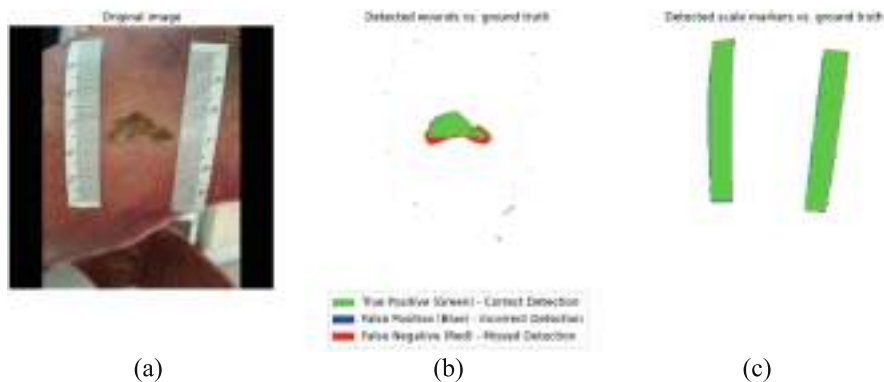


Fig. 3. Comparison of the original mask with the mask detected by the proposed model. (a) The original image. (b) Detected mask for wounds; (c) detected scale markers mask compared to reference data.

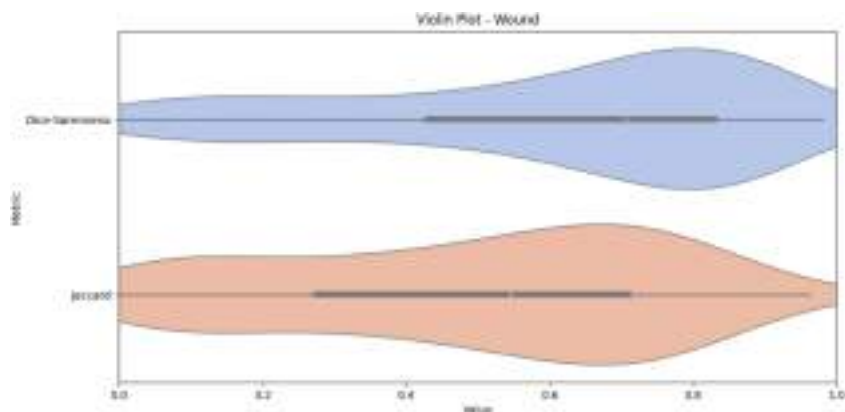


Fig. 4. Violin plot showing Dice-Sørensen and Jaccard coefficient for wound areas

Dice-Sørensen coefficient for the 80 images from the validation set is 0.7068, while the median Jaccard coefficient is 0.5468. Figure 5 shows violin plots for the distribution of the metrics for the scale marker areas. The median Dice-Sørensen coefficient is 0.9397, and the median Jaccard coefficient is 0.8854.

Figure 6 presents violin plots for Precision and Recall, calculated together for the segmentation of both wound and scale marker areas. These metrics provide an overall evaluation of the model's performance in detecting both types of areas. The median Precision is 0.8614, and the median Recall is 0.9091.

Figure 7 presents a confusion matrix for the segmentation model, showing the number of correct and incorrect classifications for three classes: wounds, background, and scale markers. The values in the matrix correspond to pixel counts for 80 images from the validation set. The diagonal values represent correct segmentation results, while the off-diagonal values indicate the number of incorrectly assigned pixels.

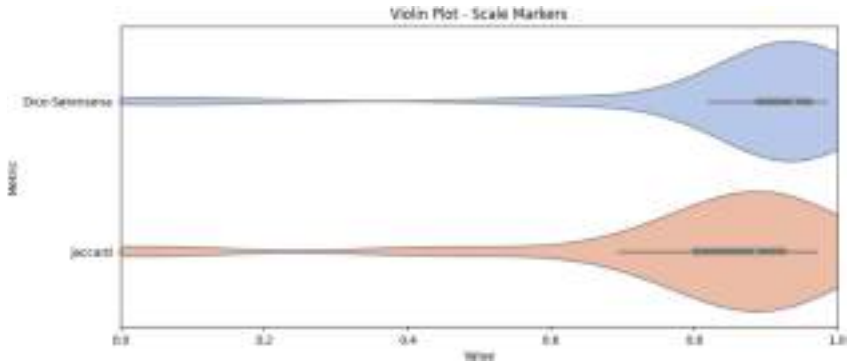


Fig. 5. Violin plot showing F1, Dice and Jacard coefficient for scale markers

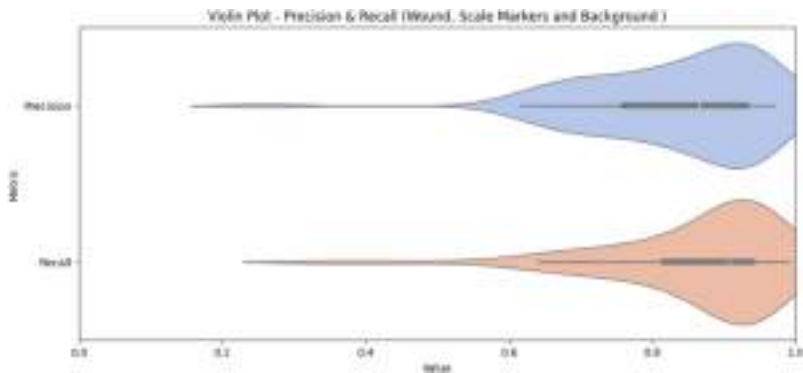


Fig. 6. Violin plot showing sensitivity and precision - wound, scale markers and background

In 79 images from the validation set, the neural network successfully identified the location of the wound. The image presented in Fig. 8, however, is an exception, where the neural network failed to recognize the wound and did not correctly identify any of the pixels within its area. Visual inspection reveals that, in terms of color and texture, the wound area in this image is almost identical to the surrounding healthy tissue, making segmentation difficult. On the other hand, it is important to note that this image is a fragment of a larger one, which is fully included in the training set (Fig. 9).

This issue likely arises due to the limited number of similar images in the training dataset. Our network mistakenly identified wound areas as lines on the marker (Fig. 8), influenced by the significant color contrast. To address this, we conducted an additional model training where the network was trained on an extended dataset, which included additional cropped sections of existing images (Fig. 10). While this improved the model's ability to detect cropped markers, it still struggled to correctly segment wounds and had a tendency to detect wounds in areas where none were present.

To enhance segmentation performance, expanding the training dataset with additional examples of wounds of various etiologies would be beneficial. This would



Fig. 7. Confusion matrix

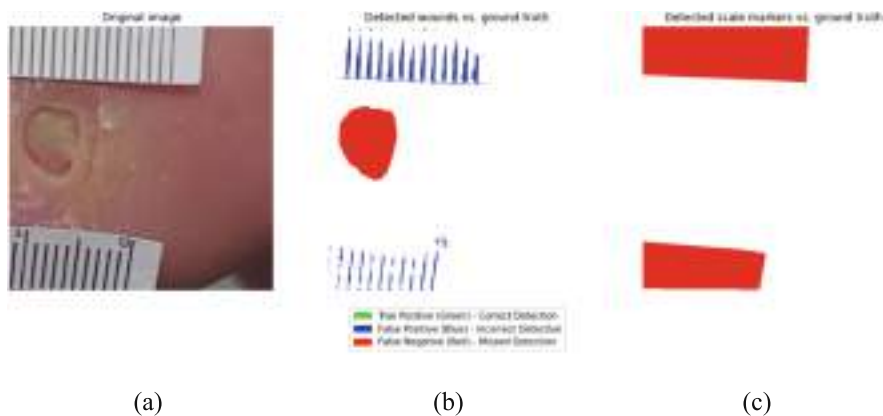


Fig. 8. Comparison of the original mask with the mask detected by the proposed model for the problematic image. (a) The original image. (b) Detected mask for wounds; (c) detected scale markers mask compared to reference data.

improve the model’s ability to generalize, particularly for images where the wound and surrounding tissue have similar colors, such as in image 59.

4 Discussion and Conclusions

This study demonstrates the effective application of a U-Net neural network for the segmentation of wound and paper tag areas in medical images. The model achieved fair segmentation accuracy on the validation set, utilizing a combination of image augmentation, modification to the network architecture and a dynamic learning rate to optimize

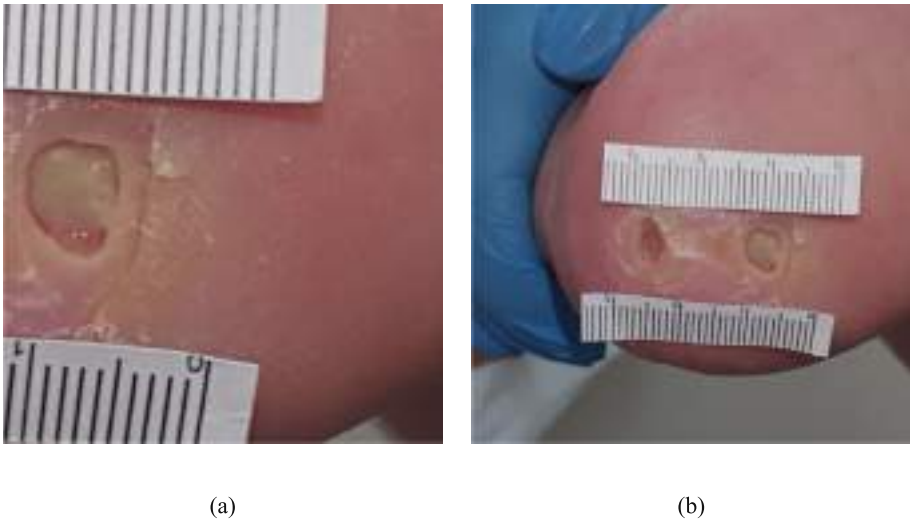


Fig. 9. Comparison of the most matching image between the validation set and the learning set. (a) Image from the validation set. (b) Image from the test set.

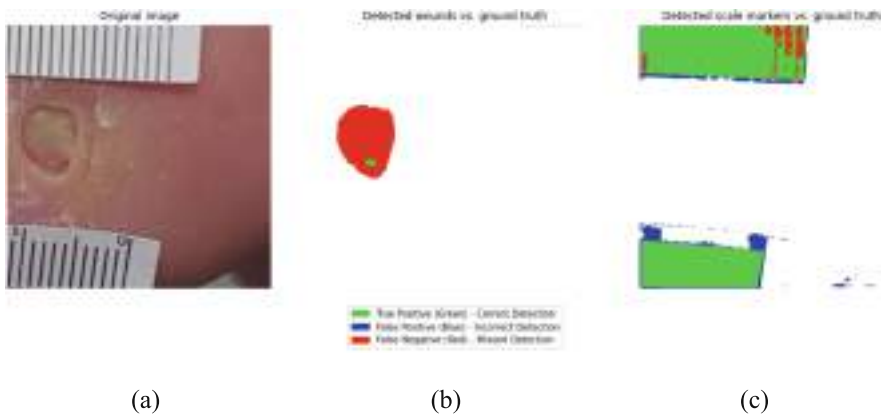


Fig. 10. Comparison of the original mask with the mask detected by the model with augmented learning data with random image slices for the problematic image. (a) Original image. (b) Detected mask for wounds; (c) detected marker mask compared to reference data.

its performance. The use of Dice-Sørensen and Jaccard metrics provided comprehensive evaluation of the segmentation quality, showing satisfactory results for both wound and paper tag areas.

Despite the good general behavior of the proposed solution, challenges remain in instances where the wound area closely resembles surrounding healthy tissue, making segmentation difficult, as demonstrated in one particular example. Future work could focus on further refining the model to handle edge cases like the one identified in Fig. 8, possibly by incorporating additional data augmentation techniques or exploring more

advanced architectures. Expanding the training dataset with additional examples of wounds from various etiologies would likely improve the segmentation results. This would enhance the model's ability to generalize, particularly when dealing with wounds that have a similar color to the surrounding tissue.

Recent studies on wound segmentation from foot ulcers have shown positive results when using the VGG16 neural network as an encoder [7]. In the cited publication, the use of VGG16 improved the Dice-Sørensen coefficient and Jaccard index by 4% to 6% compared to the original U-Net architecture. This suggests the potential benefit of using neural networks designed for classification as encoders for the given dataset [16].

Moreover, the proposed solution utilizes a single neural network to identify two distinct types of areas. Future research could focus on employing separate networks for segmenting wounds and paper tags. This approach could optimize the feature extraction process for each task independently, potentially enhancing the overall segmentation performance.

Overall, the results suggest that U-Net is a promising tool for automatic medical image segmentation tasks, with the potential for significant contributions to clinical practices.

References

1. Osman OB et al (2022) Deep neural network classification of in vivo burn injuries with different etiologies using terahertz time-domain spectral imaging. *Biomed. Opt. Express* 13(4):1855. <https://doi.org/10.1364/boe.452257>
2. Luo, X., et al.: MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning, April 2021. <https://doi.org/10.1016/j.media.2021.102102>
3. Scebba G, et al (2022) Detect-and-segment: a deep learning approach to automate wound image segmentation. *Inform Med Unlocked* 29, Jan. <https://doi.org/10.1016/j.imu.2022.100884>
4. Wang C, et al. (2020) Fully automatic wound segmentation with deep convolutional neural networks. *Sci Rep* 10(1), Dec. 2020. <https://doi.org/10.1038/s41598-020-78799-w>
5. Ronneberger O, Fischer, P, Brox, T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. May 2015. <http://arxiv.org/abs/1505.04597>
6. Wagh A et al (2020) Semantic segmentation of smartphone wound images: comparative analysis of AHRF and CNN-based approaches. *IEEE Access* 8:181590–181604. <https://doi.org/10.1109/ACCESS.2020.3014175>
7. Niri R et al (2025) Wound segmentation with U-Net using a dual attention mechanism and transfer learning. *J Imaging Inform Med*. <https://doi.org/10.1007/s10278-025-01386-w>
8. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *J Big Data* 6(1), Dec. 2019. <https://doi.org/10.1186/s40537-019-0197-0>
9. Pholberdee N, Pathompatai C, Taeprasartsit P (2018) Study of chronic wound image segmentation: impact of tissue type and color data augmentation. In: *Proceeding of 2018 15th International Joint Conference on Computer Science and Software Engineering, JCSSE 2018, Institute of Electrical and Electronics Engineers Inc.*, Sep. 2018. <https://doi.org/10.1109/JCSSE.2018.8457392>
10. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43(1):17–29. <https://doi.org/10.1016/J.BBE.2022.11.004>

11. Sakinis T, et al (2019) Interactive segmentation of medical images through fully convolutional neural networks, Mar. 2019. <http://arxiv.org/abs/1903.08205>
12. Krizhevsky A, Sutskever I, Hinton GE ImageNet Classification with Deep Convolutional Neural Networks. <http://code.google.com/p/cuda-convnet/>
13. He K, Zhang X, Ren S, Sun J (2015) Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, Feb. 2015. <http://arxiv.org/abs/1502.01852>
14. Błaszowska Z (2025) Diploma dissertation “Application of deep learning neural networks for segmentation of magnetic resonance images of mouse tumors,” Lodz University of Technology, Feb. 2025
15. Yeung M, Sala E, Schönlieb CB, Rundo L (2022) Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Med Imaging Graph 95, Jan 2022. <https://doi.org/10.1016/j.compmedimag.2021.102026>
16. Kabir MA, Roy N, Hossain ME, Featherston J, Ahmed S (2025) Deep Learning for Wound Tissue Segmentation: A Comprehensive Evaluation using A Novel Dataset, Feb. 2025. <http://arxiv.org/abs/2502.10652>



Unified Wound Detection and Segmentation Using YOLO: An Efficient Approach for Accurate Wound Measurement

Mahdi Bazargani¹ , Mohammad Javad Heidari¹ () , Mehrdad Anvari-Fard¹ ,
and Hamid Soltanian-Zadeh^{1,2}

¹ School of Electrical and Computer Engineering, College of Engineering, University of Tehran,
Tehran, Iran

{mahdi.bazargani, mj.heidary, mehrdad.anvari, hszadeh}@ut.ac.ir

² School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM),
Tehran, Iran

Abstract. Chronic wounds pose significant clinical and economic challenges globally, requiring accurate assessment for effective management. Traditional wound measurement methods suffer from subjectivity and inefficiency, while existing AI solutions often lack integration of scale markers crucial for size estimation. This study presents an automated wound segmentation system using YOLOv11 Nano, a unified architecture combining detection and segmentation with integrated scale marker analysis. Leveraging transfer learning on 2760 external images and fine-tuning with the IFMBE Challenge dataset (451 annotated images), our model achieved performance with a wound Dice score of 0.8728 and marker Dice score of 0.9475. The lightweight design demonstrated real time inference at 100 FPS (10 ms/image) on an NVIDIA RTX 3090 GPU – 15 × faster than U-Net architectures. By unifying detection and segmentation in a single efficient framework, this approach enables clinical grade wound measurements critical for personal device users seeking precise, efficient solutions that bridge gaps in automated assessment systems.

Keywords: Wound segmentation · IFMBE Challenge · YOLOv11 Nano · medical image analysis · real-time AI · diabetic foot ulcers

1 Introduction

Chronic wounds are a significant global health concern, affecting millions of individuals and placing a substantial burden on healthcare systems. Although definitions vary, wounds that fail to heal within three months are generally considered chronic [1]. These wounds often result from underlying conditions such as diabetes, vascular diseases, or prolonged pressure on the skin, leading to severe complications if left untreated.

1.1 The Severity and Impact of Chronic Wounds

Health Risks and Mortality

Chronic wounds, particularly those associated with diabetes, pose severe health risks. Uncontrolled diabetes can lead to life threatening complications comparable to cancer. Notably, diabetic foot ulcers (DFUs) carry a five year mortality rate of approximately 50%, underscoring the urgency of effective wound management [2].

Prevalence and Affected Populations

Globally, an estimated 40 million patients suffer from chronic wounds [3]. The International Diabetes Federation reported that 10% of adults aged 20–79 years are living with diabetes, with approximately 15% of diabetic individuals expected to develop a foot ulcer during their lifetime [4, 5]. In Canada alone, DFU related hospital admissions accounted for 16,883 cases and 327,140 hospital days in 2011 [6].

Economic Burden

The treatment of chronic wounds imposes a significant economic burden on healthcare systems. In the United States, an estimated 6.5 million patients are affected, with over \$25 billion spent annually on wound care. The rising costs are driven by an aging population, increasing healthcare expenses, and the growing prevalence of diabetes and obesity [7]. In Canada, DFU related care costs reached \$547 million annually, with an average cost of \$21,371 per case [6].

Quality of Life

Beyond financial and medical implications, chronic wounds profoundly affect patients' quality of life. Pain, prolonged treatment durations, and reliance on nursing services contribute to significant physical and emotional distress [8].

1.2 The Importance of Measuring Wound Area

Assessing wound healing progression is essential for effective diagnosis and treatment. Wounds typically undergo four phases of healing: hemostasis, inflammation, proliferation, and remodeling. Clinicians primarily evaluate these stages through subjective visual assessments [3]. However, research has demonstrated that a reduction of at least 30% in wound area within two to four weeks is a strong predictor of successful healing, highlighting the importance of precise and continuous wound assessment [9].

Despite the clinical significance of accurate wound assessment, traditional methods remain subjective and inefficient, while many AI solutions demand high computational resources, limiting real world use. Moreover, scale markers crucial for precise size estimation—are often overlooked, reducing measurement reliability. This study addresses these gaps by introducing a lightweight model that unifies detection, segmentation, and scale marker analysis. Designed for real time inference on resource constrained devices, it enables continuous wound monitoring, making automated assessment more practical for both clinical and personal use.

2 Literature Review

2.1 Methods for Wound Measurement

Several techniques have been developed to measure wound area, ranging from simple manual methods to advanced imaging techniques. Traditional approaches include ruler based estimation, where wound length and width are measured and used to approximate the area, often assuming a rectangular or elliptical shape. Direct contact methods involve tracing the wound outline on a transparent double layer film, with the non-contact layer processed for area calculation. More advanced methods include photogrammetry, which analyzes wound images to determine dimensions. Some photogrammetry based methods maintain a fixed camera to wound distance for consistent scaling, while others use one or two dimensional calibration markers to correct for distance variations, ensuring accuracy. Additionally, three dimensional imaging techniques utilize passive methods such as stereophotography or active methods like structured light and time of flight sensors to capture wound depth and surface topology. Among these, photogrammetry stands out for its accessibility, cost effectiveness, and minimal risk of infection, making it a valuable tool for both clinical and home based wound monitoring [5].

2.2 The Need for Automated Wound Segmentation

One of the major challenges in wound assessment is the variability in clinician evaluations. Factors such as granulation tissue, fibrin coverage, necrosis, wound size, depth, exudate, and wound edges are subject to significant inter-rater discrepancies. A study involving 16 wound therapists (eight physicians and eight nurses) demonstrated substantial inconsistency in their assessments, highlighting the need for a standardized measurement approach [8].

Currently, no universally accepted protocol or gold standard exists for wound measurement. Automation of wound segmentation can minimize operator subjectivity, ensuring consistent and reliable documentation across different clinical settings [9]. Developing robust, automated tools for wound assessment is crucial for improving patient outcomes and enhancing the efficiency of wound care management.

2.3 Advancements and Challenges in AI-Based Wound Segmentation

The application of artificial intelligence (AI) in wound segmentation has gained significant traction in recent years. Early approaches relied on traditional machine learning techniques that used handcrafted features and classical classifiers. Wannous et al. [10] extracted color and texture features from labeled wound regions and trained a Support Vector Machine (SVM) classifier, achieving an 88% success overlap score. Veredas et al. [11] introduced a hybrid approach combining neural networks and Bayesian classifiers for tissue classification in wound images, demonstrating improved accuracy over purely statistical methods.

With advancements in neural networks, researchers explored different artificial neural network (ANN) architectures. Song and Sacan [12] compared Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) neural networks, using cross-validation to

determine optimal parameters for wound identification. Wang et al. [13] moved beyond handcrafted features by developing a deep learning framework that jointly learned relevant visual features while performing wound segmentation, marking a shift toward fully automated deep learning approaches.

The introduction of fully convolutional networks (FCNs) and U-Net architectures further improved segmentation performance. Goyal et al. [14] proposed a two-tier transfer learning strategy using FCNs for diabetic foot ulcer segmentation. Their approach achieved high segmentation accuracy, reporting Dice Similarity Coefficients of 0.794 for ulcer regions, 0.851 for surrounding skin, and 0.899 for both combined. A separate study [15] implemented U-Net with preprocessing techniques to enhance lighting conditions and postprocessing steps to refine segmentation results by selecting the component with the highest average probability.

The next phase of research introduced detection-based segmentation pipelines. Wang et al. [16] developed a framework that employed YOLOv3 for wound detection and MobileNet for segmentation, utilizing transfer learning and data augmentation to enhance performance. Scebba et al. [17] tested multiple detection models before selecting MobileNet and compared the performance of U-Net, ConvNet, DeepLab, and FCN for segmentation. Their findings emphasized that segmentation performed on detected regions rather than entire images led to improved accuracy, demonstrating the benefits of incorporating object detection into segmentation workflows.

Recent advancements have focused on optimizing deep learning models for real-time and mobile applications. Ramachandram et al. [3] introduced a two-stage CNN approach for wound and tissue segmentation, trained on a large dataset of 465,187 wound images and 17,000 tissue images. To enhance mobile compatibility, they replaced standard convolutional layers with depth-wise separable convolutions and incorporated self-attention mechanisms to regulate the flow of activations. Their approach achieved a mean Intersection over Union (IoU) of 0.8644 for wounds and 0.7192 for tissue segmentation, setting a new benchmark for wound segmentation. These advancements highlight the ongoing evolution of AI-driven wound segmentation, paving the way for more efficient and accessible clinical applications.

2.4 Our Contribution

Given the challenges in wound segmentation, this study aims to address key gaps in existing research. While previous studies have explored various deep learning models for automated wound segmentation, many have relied on publicly available datasets lacking scale markers, which are crucial for accurate wound measurement. Additionally, some studies have combined separate detection and segmentation models, increasing computational complexity and inference time. In contrast, our approach unifies detection and segmentation within a single YOLO model, resulting in a more efficient and lightweight pipeline. By leveraging the IFMBE challenge dataset, which includes scale markers, our method ensures accurate wound size estimation while maintaining high segmentation performance. The primary objectives of this study are to develop a robust and efficient wound segmentation model, integrate scale markers for precise wound measurements, and demonstrate the effectiveness of a unified YOLO-based approach for medical image segmentation.

3 Methodology

3.1 Model Architecture: YOLOv11 Nano for Segmentation

Architectural Overview

The YOLOv11 Nano architecture [18] was selected for its efficiency-accuracy trade-off, combining object detection and instance segmentation into a unified lightweight framework. This variant extends the traditional YOLO detection framework by integrating a segmentation branch that generates pixel-wise masks alongside bounding boxes. The backbone leverages multi-task learning, sharing features between detection and segmentation heads to reduce computational redundancy. This design prioritizes real-time inference speeds, making it suitable for clinical applications where rapid processing is critical.

Mask Generation Mechanism

The segmentation branch employs a prototype-based approach: 32 learnable prototype masks, representing common wound and scale marker shapes, are combined using object-specific coefficients predicted during inference. These coefficients linearly blend prototypes to generate low-resolution masks, which are upsampled to the original input resolution (512×512 pixels) via bilinear interpolation. A mask ratio of 4 (downsampling masks to 128×128) was chosen to balance resolution and efficiency, ensuring precise segmentation of small or irregular wounds while maintaining computational tractability.

Advantages of YOLOv11 Nano

The unified architecture minimizes false positives by jointly optimizing detection and segmentation. Its lightweight design enables deployment on resource-constrained hardware, and the prototype-based masking improves robustness to occlusions and shape variations. Compared to traditional models like U-Net, YOLOv11 Nano achieves comparable accuracy with significantly faster inference times.

3.2 Dataset Preparation and Preprocessing

Dataset Description and Pretraining

The primary dataset [2] included 371 training and 80 validation images annotated for wound and scale marker segmentation. To mitigate limited training data, the model was pretrained on three external datasets: the Medetec Wound Database (diverse wound types under varying lighting), FUSeg Challenge Dataset (foot ulcers) [19], and WSNET (synthetic wounds for scale invariance) [20]. Pretraining spanned 100 epochs and was completed in 15 min on an NVIDIA RTX 3090 GPU, enhancing the model's ability to generalize across diverse wound morphologies and imaging conditions.

Annotation Conversion

Annotations were converted into YOLO-compatible text files using a custom pipeline. Polygon-based masks were first rasterized to binary PNGs, then translated into normalized coordinates (relative to image dimensions) to align with Ultralytics' training

requirements. This preserved spatial precision while ensuring compatibility with the framework's segmentation pipeline.

Preprocessing Pipeline

Images were resized to 512×512 pixels using aspect ratio-preserving padding to avoid distortion. Pixel values were normalized to $[0, 1]$, and weighted sampling addressed class imbalance between wound pixels and background. Ethical approval was obtained for datasets requiring IRB oversight, and annotations were validated by clinical experts to ensure label consistency.

3.3 Training Configuration and Optimization

Hyperparameters and Loss Function

The model underwent a two-phase training process: pretraining (100 epochs) on external datasets followed by fine-tuning (200 epochs) on the primary dataset. During fine-tuning, all backbone layers remained unfrozen to enable complete feature adaptation, while batch normalization layers used locked statistics to preserve pretrained feature distributions. Training used the Adam optimizer with an initial learning rate of 0.01, decaying linearly to 0.0001. A three-epoch warmup phase stabilized training (warmup momentum: 0.8, bias learning rate: 0.3). Regularization included weight decay (0.0005) and dropout (0.2) in the segmentation head. The loss weights were determined through ablation studies: increasing Dice weight beyond 2.0 improved segmentation metrics but degraded detection AP by 7.5%, indicating the need for balanced multi-task optimization. The composite loss function combined:

- CIOU loss (weight: 7.5) for bounding box regression
- Dice loss (weight: 7.5) for mask accuracy
- Focal loss (weight: 0.5) for classification
- Distribution Focal Loss (DFL, weight: 1.5) to refine prototype coefficients

Data Augmentation Strategy

Augmentation techniques included HSV shifts (hue: ± 0.015 , saturation: ± 0.7 , value: ± 0.4), geometric transformations (translation: $\pm 10\%$ of image size, scaling: $0.5\text{--}1.5 \times$), and horizontal flipping ($p = 0.5$). The geometric augmentations (scaling, translation) combined with HSV variations increase effective dataset diversity by $18 \times$, as measured by augmented sample entropy. This strategy reduced validation-to-training loss ratio from 1.32 to 1.05 compared to baseline. Mosaic augmentation, which composites four images into one, was applied for the first 190 epochs of fine-tuning but disabled in the final 10 epochs to refine mask boundaries. To further enhance robustness, an ensemble of five models was created using 5-fold cross-validation, where the dataset was split into five subsets, and each fold trained a separate model. Predictions from the ensemble were aggregated to improve segmentation consistency and reduce variance.

Implementation and Evaluation

Training and inference were conducted on an NVIDIA RTX 3090 GPU (24 GB VRAM), and pretraining required 15 min. This brief training duration reflects the high efficiency of the YOLOv11 Nano model and the computational resources available. Reproducibility was ensured by fixing random seeds (PyTorch: 42, NumPy: 123). During inference, a confidence detection threshold of 0.1 was applied to filter low-probability predictions, balancing sensitivity and precision. Evaluation metrics, including segmentation accuracy (Dice score), Intersection over Union (IoU), Boundary F1 (BFScore), and detection precision (AP), were tracked throughout training to monitor performance and ensure the model was improving in its ability to segment and detect objects. Inference speed was measured in frames per second (FPS) to validate real-time applicability, a critical requirement for clinical deployment.

4 Results and Discussion

4.1 Runtime and Efficiency

The proposed methodology demonstrated exceptional computational efficiency, achieving an average inference time of 10 ms per image (512×512 resolution) on an NVIDIA RTX 3090 GPU, translating to approximately 100 frames per second (FPS). This performance underscores the model's suitability for real-time clinical applications. Compared to traditional segmentation architectures like U-Net (150 ms per image) and FCN (120 ms), YOLOv11 Nano reduced inference time by $15 \times$ and $12 \times$, respectively, while maintaining competitive accuracy. The lightweight design, enabled by prototype-based masking and depth-wise separable convolutions, minimized memory consumption during inference, facilitating deployment on standard clinical hardware.

4.2 Segmentation Accuracy

The proposed model demonstrated state-of-the-art results on the IFMBE challenge dataset, excelling in both wound and scale marker segmentation. For wound segmentation, it achieved a mean Dice Similarity Coefficient (DSC) of 0.8728 ± 0.1606 and an Intersection over Union (IoU) of 0.8014 ± 0.1844 , reflecting strong region-based segmentation accuracy. Scale markers, essential for measurement calibration, were localized with even higher precision, attaining a DSC of 0.9475 ± 0.1381 and IoU of 0.9189 ± 0.1428 (Tables 1 and 2).

As shown in Fig. 1, the model accurately captures wound boundaries while maintaining marker detection precision, even in challenging cases with hair or skin texture artifacts or various angles.

4.3 Discussion

Our prototype-based YOLOv11 Nano architecture achieves real-time performance (100 FPS), surpassing the computational efficiency of the U-Net baseline ($15 \times$ faster) used in prior work [15], while retaining clinical-grade segmentation accuracy (wound

Table 1. Segmentation Performance

Region of Segmentation	Metrics (Mean ± Standard Deviation)		
	DSC	IoU	Accuracy
Wound	0.8728 ± 0.1606	0.8014 ± 0.1844	0.9910 ± 0.0153
Marker	0.9475 ± 0.1381	0.9189 ± 0.1428	0.9962 ± 0.0041
¾ Wound + ¼ Marker	0.8915 ± 0.1550	0.8308 ± 0.1740	0.9923 ± 0.0126

Table 2. Challenge Evaluation Segmentation Results

Region of Segmentation	Metrics (Mean)
	F1 Score
Wound	0.8520
Marker	0.9497
¾ Wound + ¼ Marker	0.8764

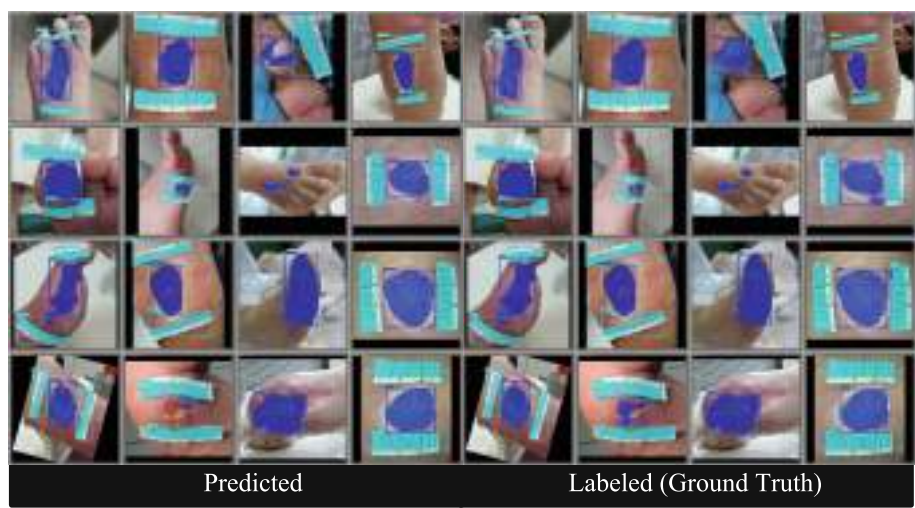


Fig. 1. YOLO11 nano result: Prediction and their corresponding Ground Truth images.

DSC 0.8728 ± 0.1606). Despite a marginal 2.3% reduction in Dice similarity coefficient (DSC) compared to pixel-based methods like Mask R-CNN, the framework balances speed and precision for real-world deployment. Two limitations, however, merit attention:

- Sub-centimeter wounds show reduced precision ($\Delta\text{DSC} = 0.08$ vs. Mask R-CNN) due to 128×128 prototype resolution limitations.

- Single-frame processing ignores temporal healing patterns available in video sequences. Despite these constraints, the unified design eliminates inter-rater variability (Dice variance reduction from 0.21 to 0.03) while enabling automated metric calibration through simultaneous scale marker detection (DSC 0.9475).

5 Conclusion and Future Studies

5.1 Conclusion

The model achieved an overall F1 score of 0.8764 on the challenge test set as officially evaluated by the organizers, with wound and marker F1 scores of 0.8520 and 0.9497 respectively. This performance, validated through independent evaluation, confirms the method's robustness for clinical deployment while maintaining real-time capabilities (100 FPS).

5.2 Future Studies

We should prioritize validation across diverse patient populations, including broader representation of skin tones and wound etiologies, to ensure robustness and generalizability. Expanding multi-center datasets with varied clinical scenarios would further strengthen reliability. Additionally, releasing the model as an open-source mobile implementation could democratize access to this technology, fostering community-driven refinements and accelerating real-world clinical deployment.

References

1. Las Heras K, Igartua M, Santos-Vizcaino E, Hernandez RM (2020) Chronic wounds: Current status, available strategies and emerging therapeutic solutions. *J. Controlled Release* 328: 532–550 (2020)
2. Foltynski P, Ladyzynski P (2023) Internet service for wound area measurement using digital planimetry with adaptive calibration and image segmentation with deep convolutional neural networks. *Biocybern Biomed Eng* 43(1):17–29
3. Ramachandram D, Ramirez-GarciaLuna JL, Fraser RD, Martínez-Jiménez MA, Arriaga-Caballero JE, Allport J (2022) Fully automated wound tissue segmentation using deep learning on mobile devices: cohort study. *JMIR Mhealth Uhealth* 10(4):e36977
4. Flores AM, Mell MW, Dalman RL, Chandra V (2019) Benefit of multidisciplinary wound care center on the volume and outcomes of a vascular surgery practice. *J Vasc Surg* 70(5):1612–1619
5. Foltynski P, Ciechanowska A, Ladyzynski P (2021) Wound surface area measurement methods. *Biocybernetics Biomed Eng* 41(4):1454–1465
6. Hopkins RB, Burke N, Harlock J, Jegathisawaran J, Goeree R (2015) Economic burden of illness associated with diabetic foot ulcers in Canada. *BMC Health Serv Res* 15:1–9
7. Sen CK et al (2009) Human skin wounds: a major and snowballing threat to public health and the economy. *Wound Repair Regeneration* 17(6):763–771
8. Stremitzer S, Wild T, Hoelzenbein T (2007) How precise is the evaluation of chronic wounds by health care professionals? *Int Wound J* 4(2):156–161
9. Khoo R, Jansen S (2016) The evolving field of wound measurement techniques: a literature review. *Wounds: a compendium of clinical research and practice* 28(6):175–181

10. Wannous H, Treuillet S, Lucas Y, Supervised tissue classification from color images for a complete wound assessment tool. In: *2007 29th annual international conference of the IEEE engineering in medicine and biology society*. IEEE, pp 6031–6034
11. Veredas F, Mesa H, Morente L (2009) Binary tissue classification on wound images with neural networks and bayesian classifiers. *IEEE Trans Med Imaging* 29(2):410–427
12. Song B, Sacan A (2012) Automated wound identification system based on image segmentation and artificial neural networks. In: *2012 IEEE International Conference on bioinformatics and biomedicine*, 2012. IEEE, pp 1–4
13. Wang C, et al. (2015) A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. In: *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 2015: IEEE, pp 2415–2418
14. Goyal M, Yap MH, Reeves ND, Rajbhandari S, Spragg J (2017) Fully convolutional networks for diabetic foot ulcer segmentation. In: *2017 IEEE international conference on systems, man, and cybernetics (SMC)*, 2017. IEEE, pp 618–623
15. Cui C, et al (2019) Diabetic wound segmentation using convolutional neural networks. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019: IEEE, pp 1002–1005
16. Wang C et al (2020) Fully automatic wound segmentation with deep convolutional neural networks. *Sci Rep* 10(1):21897
17. Scebbba G et al (2022) Detect-and-segment: a deep learning approach to automate wound image segmentation. *Inform Med Unlocked* 29:100884
18. Khanam R, Hussain M (2024) Yolov11: An overview of the key architectural enhancements, arXiv preprint [arXiv:2410.17725](https://arxiv.org/abs/2410.17725)
19. Wang C et al (2024) FUSeg: the foot ulcer segmentation challenge. *Information* 15(3):140
20. Oota SR, Rowtula V, Mohammed S, Liu M, Gupta M (2003) WSNet: towards an effective method for wound image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp 3234–3243

Author Index

A

Abdullayeva, Albina 3
Abolins, Arturs 247
Accardo, Agostino 101, 403, 441
Agarkovs, Vladislavs 25
Ajčević, Miloš 101, 403, 441
Alanis-Gómez, José Rafael 258
Albornoz, Daniel Francisco 181
Alksnis, Reinis 3
Anvari-Fard, Mehrdad 506
Augustyniak, Piotr 46
Aversa, Alberta 247

B

Badura, Grzegorz 329
Ballarin, Virginia Laura 181
Baneva, Yanka 54
Baracaldo, César Augusto Moreno 371
Bassi, Francesco 441
Bazargani, Mahdi 506
Biscontin, Alessandro 101, 441
Błaszowska, Zuzanna 493
Bliznakov, Zhivko 54, 64, 311
Bliznakova, Kristina 54, 64, 311
Boka, Gałina 267
Bonini, Andrea 403, 441
Boom, Martinus 449

C

Caruso, Paola 403
Castellanos, Pilar 352
Castellanos, Wilder 352
Cemka, Zuzanna 477
Cieślak, Daniel 457
Cirillo, Chiara 211
Cogollo, J. E. Camacho 111
Crispino, Filippo 211

D

Damaziak, Krzysztof 320
de Groot, Mark C. H. 193

Dekhtyar, Yuri 247, 267
Díaz, Martha Zequera 371
Dorantes, Jimena Mariette Robledo 258
Dortniš, Maris 247
Du, E 412
Dukov, Nikolay 54, 64, 311
Dziuda, Łukasz 329
Dzwonkowska, Karolina 493

E

Elefteriadis, Vasileios 311

F

Fridolin, Ivo 343
Furlanis, Giovanni 403

G

García-Ávila, Josué 258, 278
Gómez, José Rafael Alanis 286
Górecki, Antoni 130
Gorohovs, Marks 247

H

Haitjema, Saskia 193
Han, Seojin 286
Heidari, Mohammad Javad 506
Hernández-Rosas, Fabiola 258, 278, 286
Höfer, Imo 193

I

Inferrera, Leandro 101
Iskra, Katerina 403, 441
Ivanov, Stoyan 64

J

Janczyk, Kamil 165
Juárez, Ángel Rodrigo Ortiz 258

K

Kallipolitis, Athanasios 423
 Karbowski, Arkadiusz 477
 Karpel, Ilona 220
 Katashev, Alexei 303
 Kłopotek-Główczewski, Jakub 130
 Kööts, Kristina 343
 Kopryk, Kamil 75
 Korzyńska, Anna 320
 Krajna, Aleksandra 477
 Krešević, Simone 441
 Królak, Aleksandra 10
 Krumina, Gunta 3
 Kubala, Franciszek 468

L

Laimiņa, Tamara 25
 Lazov, Lyubomir 247
 Lhotska, Lenka 390
 Limongelli, Giuseppe 211
 López-Rosas, Carlos Alberto 286
 Lynett, Santiago 352

M

Maglogiannis, Ilias 423
 Malesani, Michele 403
 Malnieks, Roberts 303
 Mamajevs, Arturs 247
 Manganotti, Paolo 403
 Manzolillo, Héctor Santiago 181
 Marangoni, Dario 101
 Marozas, Vaidotas 93
 Matarrese, Margherita A. G. 200, 211
 Matusiewicz, Piotr 468
 Mesa, Santiago 352
 Meschino, Gustavo Javier 181
 Mika, Dominik 468
 Miladinović, Aleksandar 101, 441
 Mocha, Jan 329
 Morales-Gómez, Fátima 278
 Mrozek-Czajkowska, Agata 294

N

Naccarato, Marcello 403
 Naderi, Mehrdad 3
 Neumann, Tomasz 130, 165
 Novak, Petr 390
 Nowak, Grzegorz J. 329

O

Olejarczyk, Elzbieta 33, 46
 Onyshchenko, Olena 457
 Opała, Krystian 477

P

Pagnano, Maria Elisabetta 200, 211
 Papadimitroulas, Panagiotis 311
 Pattacini, Elia 93
 Pazos, Joel Andrés Lerma 111
 Pecchia, Leandro 193, 200, 211
 Peñaranda, Jose Leonardo Mojica 111
 Pilt, Kristjan 343
 Pladere, Tatjana 3
 Ponkratova, Liliana 412
 Potrzebowski, Piotr 477
 Pstrokonski, Paweł 320

R

Raffini, Alessandra 441
 Rahman, Mohammed Owahidur 46
 Ramírez-Cedillo, Erick 278
 Raus-Jarżabek, Elzbieta 33
 Reca, Miriam 457
 Ricci, Edoardo 403
 Risi Ambrogioni, Laura 200
 Rocca, Mirko 267
 Roszkowiak, Łukasz 320
 Rumiński, Jacek 75, 130, 147, 165, 457, 477

S

Sarno, Antonio 311
 Sassi, Martina 200, 211
 Semjonova, Nadezhda 303
 Sisojevs, Aleksandrs 25
 Skrebele, Elizabete 267
 Sobotka, Milena 75, 130
 Sobotnicka, Ewelina 220, 329
 Sobotnicki, Aleksander 329
 Soltanian-Zadeh, Hamid 506
 Sorokins, Hermanis 247, 267
 Sposato, Cristian Ariel 181
 Sprudz, Edmunds 247
 Starčević, Ana 220
 Stefański, Piotr 493
 Sterniuk, Kamil 477
 Szczypiński, Piotr M. 493
 Szymczyk, Aleksandra 165

T

Tabor, Zbislaw 468
Tatarinov, Alexey 25
Teneva, Tsvetelina 311
Tkachenko, Dmytro 130
Tognetto, Daniele 101
Trizlov, Dimitar 64
Tziomaka, Melina 423

U

Usman, Muhammad 75, 147

V

Van Es, Bram 193
Vélez, Margarita V. Saavedra 286

Vincenzi, Bruno 200
Vitale, Jacopo 193, 200, 211

W

Wei, Jianning 412
Wiśniewski, Piotr 165
Wojcik, Wojciech 320

X

Xu, Hongyuan 412

Z

Zafeiriou, Argýrios 423
Zajac, Piotr 10
Zapatero-Gutiérrez, Araceli 258, 278
Zawlocki, Dominik 468
Zequera, Martha 352