


**Communications
in Computer and Information Science**

2273

Editorial Board Member

Ashish Ghosh , *Indian Statistical Institute, Kolkata, India*

Rationale

The CCIS series is devoted to the publication of proceedings of computer science conferences. Its aim is to efficiently disseminate original research results in informatics in printed and electronic form. While the focus is on publication of peer-reviewed full papers presenting mature work, inclusion of reviewed short papers reporting on work in progress is welcome, too. Besides globally relevant meetings with internationally representative program committees guaranteeing a strict peer-reviewing and paper selection process, conferences run by societies or of high regional or national relevance are also considered for publication.

Topics

The topical scope of CCIS spans the entire spectrum of informatics ranging from foundational topics in the theory of computing to information and communications science and technology and a broad variety of interdisciplinary application fields.

Information for Volume Editors and Authors

Publication in CCIS is free of charge. No royalties are paid, however, we offer registered conference participants temporary free access to the online version of the conference proceedings on SpringerLink (<http://link.springer.com>) by means of an http referrer from the conference website and/or a number of complimentary printed copies, as specified in the official acceptance email of the event.

CCIS proceedings can be published in time for distribution at conferences or as post-proceedings, and delivered in the form of printed books and/or electronically as USBs and/or e-content licenses for accessing proceedings at SpringerLink. Furthermore, CCIS proceedings are included in the CCIS electronic book series hosted in the SpringerLink digital library at <https://link.springer.com/bookseries/7899>. Conferences publishing in CCIS are allowed to use Online Conference Service (OCS) for managing the whole proceedings lifecycle (from submission and reviewing to preparing for publication) free of charge.

Publication process

The language of publication is exclusively English. Authors publishing in CCIS have to sign the Springer CCIS copyright transfer form, however, they are free to use their material published in CCIS for substantially changed, more elaborate subsequent publications elsewhere. For the preparation of the camera-ready papers/files, authors have to strictly adhere to the Springer CCIS Authors' Instructions and are strongly encouraged to use the CCIS LaTeX style files or templates.

Abstracting/Indexing

CCIS is abstracted/indexed in DBLP, Google Scholar, EI-Compendex, Mathematical Reviews, SCImago, Scopus. CCIS volumes are also submitted for the inclusion in ISI Proceedings.

How to start

To start the evaluation of your proposal for inclusion in the CCIS series, please send an e-mail to ccis@springer.com.


Santiago Berrezueta-Guzman · Rommel Torres ·
Jorge Luis Zambrano-Martínez ·
Jorge Herrera-Tapia
Editors

Information and Communication Technologies

12th Ecuadorian Conference, TICEC 2024,
Loja, Ecuador, October 16–18, 2024,
Proceedings

Editors

Santiago Berrezueta-Guzman 
Technical University of Munich
Heilbronn, Germany

Jorge Luis Zambrano-Martínez 
Universidad del Azuay
Cuenca, Ecuador

Rommel Torres 
Universidad Técnica Particular de Loja
Loja, Ecuador

Jorge Herrera-Tapia 
Universidad Laica Eloy Alfaro de Manabí
Manta, Ecuador

ISSN 1865-0929

ISSN 1865-0937 (electronic)

Communications in Computer and Information Science

ISBN 978-3-031-75430-2

ISBN 978-3-031-75431-9 (eBook)

<https://doi.org/10.1007/978-3-031-75431-9>

© 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

Welcome to the twelfth edition of the Ecuadorian Conference on Information and Communication Technologies (TICEC 2024). This year, the conference was held in the historic city of Loja from October 16 to 18, 2024, at the Universidad Técnica Particular de Loja (UTPL).

TICEC is an important event in Ecuador focused on the development of Information and Communication Technologies (ICTs). Every year, it brings together researchers, educators, professionals, and students from around the world to share their research and engage in academic discussions. These interactions aim to promote innovative uses of ICTs across different fields.

This year, the event's organization was masterfully executed through a partnership between the Universidad Técnica Particular de Loja and the Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (CEDIA). The conference featured insightful oral presentations of scientific papers spanning three primary domains:

- Image Processing, Classification, and Segmentation.
- Artificial Intelligence and Machine Learning Applications.
- IoT, Embedded Systems, and Applications in Healthcare and Industrial Environments.

This edition witnessed submissions from 233 authors hailing from 6 countries, a testament to its growing global influence. Ensuring the integrity and quality of our proceedings, all manuscripts underwent rigorous similarity checks and were subjected to a meticulous double-blind peer review by our esteemed TICEC 2024 Program Committee. This committee comprised renowned researchers from 40 countries, with each manuscript evaluated by a minimum of three experts. From this rigorous process, we are proud to present 24 accepted full papers out of 74 received papers, marking an acceptance rate of 32%.

We are excited to welcome you to the proceedings of TICEC 2024.

October 2024

Santiago Berrezueta-Guzman
Rommel Torres
Jorge Luis Zambrano-Martínez
Jorge Herrera-Tapia

Organization

Honorary Committee

Cecilia Paredes	CEDIA, Ecuador
Santiago Acosta	Universidad Técnica Particular de Loja, Ecuador
Juan Pablo Carvallo Vega	CEDIA, Ecuador

General Chair

Rommel Torres	Universidad Técnica Particular de Loja, Ecuador
---------------	---

Steering Committee Chairs

Germania Rodriguez Morales	Universidad Técnica Particular de Loja, Ecuador
Efrain R. Fonseca C.	Universidad de las Fuerzas Armadas, Ecuador
Jorge Maldonado-Mahauad	Universidad de Cuenca, Ecuador
Marcos Orellana	Universidad del Azuay, Ecuador
Juan Pablo Salgado	Universidad Politécnica Salesiana, Ecuador
Jorge Herrera-Tapia	Universidad Laica Eloy Alfaro de Manabí, Ecuador
Jorge Luis Zambrano-Martínez	Universidad del Azuay, Ecuador
Santiago Berrezueta-Guzman	Technical University of Munich, Germany

Organizing Committee

Germania Rodriguez Morales	Universidad Técnica Particular de Loja, Ecuador
Alexandra Gonzalez Eras	Universidad Técnica Particular de Loja, Ecuador
Karina Chamba Carrillo	Universidad Técnica Particular de Loja, Ecuador
Karina Castillo Tacuya	Universidad Técnica Particular de Loja, Ecuador
Armando Cabrera Silva	Universidad Técnica Particular de Loja, Ecuador
Laura Malache-Silva	CEDIA, Ecuador
Micaela Ferrari	CEDIA, Ecuador
Francisco Toral	CEDIA, Ecuador
Santiago Ruilova	CEDIA, Ecuador
Samantha Romero	CEDIA, Ecuador

Neida Astudillo CEDIA, Ecuador
Carlos Castillo CEDIA, Ecuador

Program Committee

Alvaro Llaría	University of Bordeaux, France
Alvaro Suarez	Universidad de Las Palmas de G.C., Spain
Belen Bermejo	University of the Balearic Islands, Spain
Carlos Abreu	Instituto Politécnico de Viana do Castelo, Portugal
Christos Antonopoulos	University of the Peloponnese, Greece
Claudia Marzi	Italian National Research Council, Italy
Cristian Vasar	Politehnica University of Timisoara, Romania
Dan Pescaru	Universitatea Politehnica din Timisoara, Romania
David Valiente	Miguel Hernandez University, Spain
Efraín R. Fonseca C.	Universidad de las Fuerzas Armadas, Ecuador
Enrique Hernández-Orallo	Universidad Politécnica de Valencia, Spain
Francisco J. Martínez	University of Zaragoza, Spain
George Adam	University of Thessaly, Greece
Giuseppe Ciaburro	Università della Campania Luigi Vanvitelli, Italy
Hugo Ferreira	Polytechnic Institute of Oporto, Portugal
Iosif Szeidert	Politehnica University of Timisoara, Romania
Irina Georgiana Mocanu	University Politehnica of Bucharest, Romania
Ivan Kuzmin	Technical University of Munich, Germany
Ivan Parmacli	Technical University of Munich, Germany
Ivan Pau	Universidad Politécnica de Madrid, Spain
John Castro	Universidad de Atacama, Chile
Jorge Herrera-Tapia	Universidad Laica Eloy Alfaro de Manabí, Ecuador
Jorge Luis Zambrano-Martínez	Universidad del Azuay, Ecuador
Kerstin Andree	Technical University of Munich, Germany
Kun Qin	Technical University of Munich, Germany
Leon Bein	Technical University of Munich, Germany
Leticia Guarino	Universidad Simón Bolívar, Venezuela
Loredana Stanciu	Politehnica University of Timisoara, Romania
Lucian Pislaru-Danescu	INC DIE ICPE-CA, Romania
Lukasz Sobaszek	Lublin University of Technology, Poland
Marcelo Errecalde	Universidad Nacional de San Luis, Argentina
Marcin Górski	Silesian University of Technology, Poland
Marco Zappatore	University of Salento, Italy
María Dolón-Poza	Universidad Politécnica de Madrid, Spain
Maria Francesca Bruno	Politecnico di Bari, Italy

María Cristina Rodríguez	Rey Juan Carlos University, Spain
María-Luisa Martín-Ruíz	Universidad Politécnica de Madrid, Spain
Marian Wysocki	Rzeszow University of Technology, Poland
Marlon Navia	Universidad Técnica de Manabí, Ecuador
Melissa Montalvo	Screen Point Medical, The Netherlands
Mihail Atanasov	Technical University of Munich, Germany
Mohammad Kasra Habib	University of Stuttgart, Germany
Pietro Manzoni	Universidad Politécnica de Valencia, Spain
Przemysław Mazurek	West Pomeranian University of Technology, Poland
Shaibal Barua	Mälardalen University, Sweden
Santiago Berrezueta-Guzman	Technical University of Munich, Germany
Tomasz Bieniek	Institute of Electron Technology, Poland
Vera Ferreira	Federal University of the Pampa, Brazil
Vinayak Elangovan	Penn State Abington, USA
Vladimir Robles-Bykbaev	Universidad Politécnica Salesiana, Ecuador
Wen-Chun Chen	Technical University of Munich, Germany
Wojciech Zabierowski	Łódź University of Technology, Poland
Xavier Franch	Universitat Politècnica de Catalunya, Spain
Yanhua Luo	University of New South Wales, Australia
Yu Huang	Chinese Academy of Sciences, China
Zoltán Ádám Tamus	Budapest University of Technology, Hungary

Contents

Image Processing, Classification, and Segmentation

Image Classification of Peach Leaves Using a Siamese Neural Network	3
<i>Mateo David Coello-Andrade, Iván Reyes-Chacón, Paulina Vizcaino-Imacaña, and Manuel Eugenio Morocho-Cayamcela</i>	
Pancreas Segmentation Using SRGAN Combined with U-Net Neural Network	21
<i>Mayra Elizabeth Tualombo, Iván Reyes, Paulina Vizcaino-Imacaña, and Manuel Eugenio Morocho-Cayamcela</i>	
Using Artificial Intelligence and X-ray Images to Train and Predict COVID-19 and Pneumonia: Tool for Diagnosis and Treatment	39
<i>Bryan Juárez-Gonzalez, Fernando Villalba-Meneses, Jonathan Cruz-Varela, Andrés Tirado-Espín, Paulina Vizcaino-Imacaña, Carolina Cadena-Morejon, Cesar Guevara, and Diego Almeida-Galárraga</i>	
Region of Interest Features and Classification of MRI Brain Lesions	57
<i>Darwin Castillo, Ricardo J. Alejandro, Santiago García, María José Rodríguez-Álvarez, and Vasudevan Lakshminarayanan</i>	
Deep Learning-Based Leukemia Diagnosis from Bone Marrow Images	71
<i>Luis Zhinin-Vera, Alejandro Moya, Elena Pretel, Jaime Astudillo, and Javier Jiménez-Ruescas</i>	
Breast Thermographic Image Augmentation Using Generative Adversarial Networks (GANs)	86
<i>Ramiro Israel Vivanco Gualán, Yuliana del Cisne Jiménez Gaona, Darwin Patricio Castillo Malla, María José Rodríguez-Alvarez, and Vasudevan Lakshminarayanan</i>	
Evaluating Histopathological Cancer Detection: A Comparative Analysis of CNN Architectures for Tumor Detection in Lymph Node Pathology	100
<i>Ana Marcillo-Vera, Karen Cáceres-Benítez, Diego Almeida-Galárraga, and Andrés Tirado-Espín</i>	

ViTSigat: Early Black Sigatoka Detection in Banana Plants Using Vision Transformer 117
Jorge L. Charco, Angela Yanza-Montalvan, Johanna Zumba-Gamboa, Jose Alonso-Anguizaca, and Edgar Basurto-Cruz

Automatic Parking Space Segmentation Using K-Means Clustering and Image Processing Techniques 131

AQI

Anthony Xavier Romero Gonzalez, Kevin Sebastian Campoverde Ambrosi, Patricio Eduardo Ramon Celi, Alexandra Bermeo, Marcos Orellana, Jorge Luis Zambrano-Martinez, and Patricio Santiago García-Montero

Artificial Intelligence and Machine Learning Applications

Neural Agents with Continual Learning Capacities 145
Luis Zhinin-Vera, Elena Pretel, Alejandro Moya, Javier Jiménez-Ruescas, and Jaime Astudillo

Centinel: An Intelligent System Based on an Integrated Architecture for Supporting Scholars 160
Lorena Recalde, Gabriela Suntaxi, Diana Martínez-Mosquera, Rommel Masabanda, and Danny Cabrera

Optimizing Predictive Models in Healthcare Using Artificial Intelligence: A Comprehensive Approach with a COVID-19 Case Study 178
Juan Pablo Astudillo León, Kevin Chamorro, and Santiago J. Ballaz

Predictive Model Proposal in Telemetry Using Machine Learning Techniques to Anticipate Water Degradation in Aquaculture 193
Néstor Rafael Salinas-Buestán, Francisco Alexander Zambrano-Varela, Ángel Iván Torres-Quijije, Diego Fernando Intriago-Rodríguez, and Diego Patricio Peña-Banegas

Evaluation of the Use of Artificial Intelligence Techniques in the Mitigation of the Broadcast Storm Problem in FANET Networks 208
Andrés Sánchez, Patricia Ludeña-González, and Katty Rohoden

Adaptation Dynamics of Galápagos Finches: Evolutionary Responses to Climate Variation Explored through Machine Learning 223
Ariana Deyaneira Jiménez Narváez, Dánely Leonor Sánchez Vera, Iván Reyes, Paulina Vizcaíno-Imacaña, and Manuel Eugenio Morocho-Cayamcela

Incident Alert Priority Levels Classification in Command and Control Centre Using Word Embedding Techniques 238
Marcos Orellana, Jonnathan Emmanuel Cubero Lupercio, Juan Fernando Lima, Patricio Santiago García-Montero, and Jorge Luis Zambrano-Martinez

Physics Informed Neural Networks and Gaussian Processes-Hamiltonian Monte Carlo to Solve Ordinary Differential Equations 253
Roberth Chachalo, Jaime Astudillo, Saba Infante, and Israel Pineda

IoT, Embedded Systems, and Applications in Healthcare and Industrial Environments

Pre-processing of the Text of ECU 911 Emergency Calls 271
Marcos Orellana, Pablo Andres Molina Pinos, Patricio Santiago García-Montero, and Jorge Luis Zambrano-Martinez

Secure BLE Communication Between Android Devices and Embedded Systems for IoT Applications 285
Raúl Armas and Marlon Navia

Modeling and Control of a Peltier Thermoelectric System Applying a Multi-objective Optimization Approach 303
Víctor Huilcapi, Geovanny García, Elias Ghia, and Brian Soto

Wearable Device for Acquiring Biomechanical Variables Applied to the Analysis of Occupational Health Risks in Industrial Environments 318
Carlos Calderon-Cordova, Victor Puchaicela, and Roger Sarango

Spectral Analysis of Powertrain Vibration in a Hybrid Vehicle Under Controlled Operating Conditions 333
Raquel de los A. Salas Ibarra, Alexander E. Torres Romero, and David H. Cárdenas Villacrés

Evaluation of an Intelligent Educational Toy-Game Prototype for Toddlers’ Motor Stimulation and Learning 345
Nayeth Idalid Solorzano Alcivar, Da Hee Park Kim, Jimmy Ernesto Canizares Pozo, Michael Xavier Arce Sierra, and Andrea Paola Rubio Zurita

Changes in Academic Assessment Due to the Use of Artificial Intelligence 364
Isaac Ojeda and Santiago Castro Arias

Author Index 381

Author Queries



Chapter

Query Refs.	Details Required	Author's response
AQ1	Kindly note discrepancy found in author names between copyright page and chapter as 'Jorge Luis Zambrano-Martínez, Jorge Luis Zambrano-Martinez'. Please check and confirm.	

Image Processing, Classification, and Segmentation



Image Classification of Peach Leaves Using a Siamese Neural Network

Mateo David Coello-Andrade¹ , Iván Reyes-Chacón² ,
Paulina Vizcaino-Imacaña² ,
and Manuel Eugenio Morocho-Cayamcela^{1,2}  

¹ Yachay Tech University, School of Mathematical and Computational Sciences, DeepARC Research Group, Hda. San José s/n y Proyecto Yachay, Urcuquí 100119, Ecuador

`mateo.coello@yachaytech.edu.ec`, `mmorocho@yachaytech.edu.ec`

² Universidad Internacional del Ecuador, Faculty of Technical Sciences, School of Computer Science, Quito 170411, Ecuador
{`ireyes,pvizcaino,mamorochoca`}@uide.edu.ec

Abstract. The growing global population and the increasing demand for food have made food production a critical concern. To meet this challenge, techniques like vertical farming and aquaponics have been proposed to maximize output while conserving resources and space. However, there is still room for improvement in crop care processes. Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a leading approach for addressing various agricultural issues. This paper explores the use of Siamese CNNs to classify the front and back faces of six peach leaf varieties, a crucial step in detecting bacterial diseases, a common disease in peach farming. Building on prior work in Siamese CNNs for peach leaf classification, this study examines several enhancements, including increased convolutional layers, three activation functions, and the application of pre-trained CNN models. The proposed architecture, combined with the ReLU activation function, improved the accuracy of the reference model by 3.48%. Among the pre-trained models, ResNet performed best, reaching an accuracy of 0.9841, which is 7.83% higher than the benchmark model's top result. These findings indicate that the benchmark model had significant room for improvement, which was effectively addressed through the experiments conducted.

Keywords: Computer Vision · Image Classification · Peach Leaves Classification · Siamese Convolutional Neural Networks

1 Introduction

Food production and its impact on today's population are of paramount importance. As stated by the Food and Agriculture Organization [23], the constant growth of the world's population requires more ample food supplies. The United Nations [29] estimated that the world population will reach 9.8 billion by 2050.

Techniques such as vertical crops [6] and aquaponics [4] have been proposed to produce more food while consuming fewer resources and less space. Even though agricultural techniques are constantly improving and studies in this field are increasing, there is a need to enhance the processes of crop care [7, 11, 22].

Prior to introducing one of the modern solutions for crop caring, a general idea of image classification is given. Image classification as defined by Rawat and Wang [24] is a computer vision task aimed at categorizing images into one or multiple pre-established classes. Several are the machine-learning algorithms for image classification, such as support vector machines, k-nearest neighbors, decision trees, or random forests [18, 21, 25]. Despite machine-learning algorithms being an efficient solution, the current paradigm is focused on deep-learning methods, specifically convolutional neural networks (CNNs).

As explained by Zhong et al. [31] and Takahashi et al. [28], CNNs combine three main architectures to improve learning accuracy and efficiency, local receptive fields, shared weights, and spatial pooling. In image classification tasks, the local receptive fields are used to extract features of an image (edges and corners) that are merged with subsequent layers to build complex features. The large size of images fed into the network causes the computational complexity to rise due to the multiple layers. By implementing an architecture that handles shared weights across the layers, the number of parameters is reduced. Finally, adding a pooling architecture enables reducing the resolution of feature maps, thus, making the network invariant to small shifts in the input image. Therefore, the use of CNNs in agriculture has seen many applications due to their capacity to recognize the different features of crops and being able to detect and identify at time [16].

On the other hand, disease detection plays an important role in crop care since late recognition can cause the product to get wasted or even infect surrounding crops. Akbar et al. [2] explain that peach, a fruit high in antioxidants, vitamins, and minerals, constitutes one of the essential products of Pakistani agriculture. Nevertheless, peach production tends to be largely affected by bacteriosis that causes sized holes in leaves, rough areas on fruit, and concentric lesions on branches [14]. Acknowledging that bacteriosis commonly starts on the leaves of peach trees, allows for the implementation of a CNN model to classify infected leaves from healthy ones. However, an initial image classifier should be implemented to distinguish between the multiple kinds of peach varieties by recognizing the features of its leaves.

In this regard, this paper proposes an improved Siamese CNN to identify both the front and back faces of peach leaves. This article takes into consideration the work developed by Ayala et al. [3] whose Siamese architecture was considered the benchmark model to establish a point of comparison. This work analyzes several aspects, including increased convolutional layers, testing of three activation functions, and use of pre-trained CNN models. The experiments developed showed outperforming results in comparison to the benchmark architecture.

2 Related Works

With the advent of deep learning, convolutional neural networks (CNNs) have emerged as the state-of-the-art approach for image classification in agriculture [20]. CNNs are capable of automatically learning feature representations from raw image data, allowing for high accuracy and robustness in image classification tasks [26]. The use of CNNs for image classification in agriculture has been extensively studied, with numerous works demonstrating their effectiveness for a variety of tasks, such as crop type classification [15], plant classification [13], leaf disease detection [1], fruit maturity [17], or fruit counting [12].

Recently, there has been a trend towards fine-tuning pre-trained CNN models, such as ResNet, VGG, and Inception, for specific agriculture tasks. For instance, Subramanian et al. [27] employed pre-trained models for disease identification in maize leaves, whereas Espejo et al. [9] proposed the use of pre-trained models for weeds identification. The results of those works have demonstrated the effectiveness of transfer learning in the context of agricultural image classification. The use of fine-tuned CNNs is especially effective for small agriculture datasets, where the limited amount of data may not be sufficient to train a CNN [10] efficiently.

The integration of aerial and satellite imagery has opened up new avenues for large-scale, automated image classification in agriculture [8]. These images provide high-resolution information on crop growth and health, enabling the development of precision agriculture methods for improved crop management. The use of remote sensing for image classification in agriculture has been extensively studied, with numerous works demonstrating its effectiveness for a variety of tasks, such as crop yield estimation [19] and water stress detection [30].

In this context, current advances in image classification for agriculture offer various solutions for crop care. This work contributes to the field by focusing specifically on peach leaf image classification as a preliminary step in identifying diseases that begin developing in the leaves.

3 System Model and Methodology

3.1 Dataset

As this is a comparative study to improve the accuracy reached of the Siamese CNN proposed by Ayala et al. [3], the same dataset of the study is taken into account (see Appendix A to obtain the dataset). The dataset comprises both the front and back faces of leaves for the 6 different varieties of peach. The dataset comprises a total of 1265 pairs of images subdivided into 1139 pairs for training and 126 pairs for testing. A dataset summary is presented in Table 1, and images of peach leaf pairs for each variety of the dataset are presented in Fig. 1. The dataset has two derivatives regarding the dimensions of images, 224×224 and 416×416 , both RGB images. In this work, only the 224×224 derivative was considered.

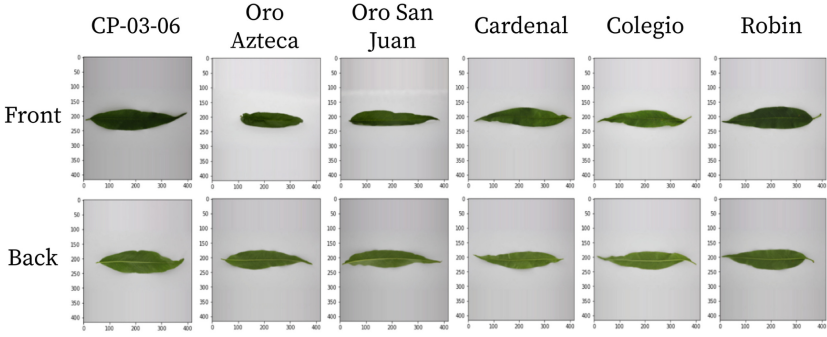


Fig. 1. Front and back faces of the six peach leaf varieties. Adapted from [3].

Table 1. Distribution of peach leaf images across the 6 varieties.

Class	Name	Train pairs	Test pairs
1	CP-03-06	90	11
2	Oro Azteca	186	21
3	Oro San Juan	179	25
4	Cardenal	209	20
5	Colegio	190	15
6	Robin	285	34

3.2 Pipeline

Figure 2 presents the pipeline of the Siamese network for identifying the front and back faces of peach leaves. The model receives two inputs, the front and back faces of a peach leaf. These are then fed into two identical convolutional neural networks. The feature maps generated at the end are concatenated and passed to a second convolutional neural network to extract more complex features. Finally, the resultant feature maps are flattened and fed to a fully connected network that will classify the leaves into their corresponding peach variety.

Table 2. Units of the benchmark CNN architecture.

(a) Type I	(b) Type II
Convolutional Layer $(K=5, P=1, S=1)$	Convolutional Layer $(K=5, P=1, S=1)$
Batch Normalization	Batch Normalization
Activation Function (ReLU)	Activation Function (ReLU)
Pooling Layer $(K=3, P=1, S=1)$	Pooling Layer $(K=2, P=1, S=1)$
	Dropout Layer $(0.05, 0.1)$

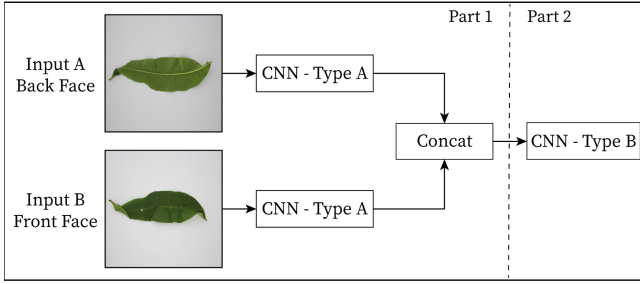


Fig. 2. Pipeline to identify peach leaves front and back faces. Adapted from [3]

3.3 Benchmark Architecture

Additionally, the Siamese CNN architecture described in [3] for the 224×224 derivative of the dataset was replicated from scratch in TensorFlow. The intention of replicating this architecture is to compare the changes introduced in the proposed architecture and the different experiments developed. The architecture consists of two units, both comprised of a convolutional layer, a batch normalization layer, an activation layer, and a max pooling layer. The activation function selected by the authors was the ReLU activation function, which is formally introduced in the following section. However, the second unit also contains a dropout layer, a representation of such units is presented in Table 2. This replication process ensures that any observed differences in performance are attributable solely to the modifications applied in the experiments. The benchmark type A CNN contains three type I units, whereas the type B CNN is comprised of two type II units. Moreover, the layers regarding the type A CNN are presented in Table 3, whereas the layers for the type B CNN are presented in Table 4.

Table 3. Layers corresponding to the type A CNN - Benchmark Paper.

Layer	Input size	Output size
Conv2d ($f = 18, k = 5, p = 1, s = 1$)	$224 \times 224 \times 3$	$224 \times 224 \times 18$
Maxpool ($k = 3, p = 1, s = 1$)	$224 \times 224 \times 18$	$75 \times 75 \times 18$
Conv2d ($f = 30, k = 5, p = 1, s = 1$)	$75 \times 75 \times 18$	$75 \times 75 \times 30$
Maxpool ($k = 3, p = 1, s = 1$)	$75 \times 75 \times 30$	$25 \times 25 \times 30$
Conv2d ($f = 60, k = 5, p = 1, s = 1$)	$25 \times 25 \times 30$	$25 \times 25 \times 60$
Maxpool ($k = 3, p = 1, s = 1$)	$25 \times 25 \times 60$	$9 \times 9 \times 60$

Table 4. Layers corresponding to the type B CNN - Benchmark Paper.

Layer	Input size	Output size
Conv2d (f = 100, k = 5, p = 1, s = 1)	$9 \times 9 \times 120$	$9 \times 9 \times 100$
Maxpool (k = 2, p = 1, s = 1)	$9 \times 9 \times 100$	$5 \times 5 \times 100$
Conv2d (f = 90, k = 5, p = 1, s = 1)	$5 \times 5 \times 100$	$5 \times 5 \times 90$
Maxpool (k = 2, p = 1, s = 1)	$5 \times 5 \times 90$	$3 \times 3 \times 90$
Flatten	$3 \times 3 \times 90$	810
Linear	810	720
Linear	720	7

3.4 Experiments

The experiments conducted to improve accuracy and reduce loss on the validation set are presented in the following sections. Likewise, each implemented model is available at the repository link provided in Appendix B. The hyperparameters selected to train the benchmark and proposed architectures were a batch size of 32, 100 epochs, and a learning rate of 0.001. In contrast, the hyperparameters chosen for experiments with pre-trained models were a batch size of 32, 50 epochs, and a learning rate of 0.0001. Categorical cross-entropy loss and the Adam optimizer were used to train the models. Due to the non-deterministic nature of a neural network’s training process, accuracy results may vary with each training session. Therefore, each model will be trained five times to gather different accuracy outcomes and provide a more comprehensive analysis. The term “iteration” will be used to refer to each individual training phase of a model.

1. Increase of convolutional layers

The first experiment involved increasing the number of convolutional layers and reorganizing the pooling layers. The Type A CNN (Table 5) consists of six convolutional and three pooling layers, while the Type B CNN (Table 6) has three convolutional and three pooling layers. No additional hidden layers were added to the fully connected network, but the first hidden layer’s neurons were set to 1000. Applying the `to_categorical()` function to the training and validation labels resulted in tensors with shapes of (1139,7) and (126,7), where the seventh category, unrelated to any peach variety, ensures the binary class matrix has full rank. After each convolutional layer, a batch normalization layer and an activation layer were added. Likewise, dropout layers were included to reduce overfitting in the training set, with a dropout of 0.1 applied after the activation layers in each unit of the type B CNN. The final designs for both CNN architectures are shown in Table 7, with type A containing three type I units and type B consisting of three type II units.

Table 5. Layers corresponding to the type A CNN - Proposed architecture.

Layer	Input Size	Output Size
Conv2d($f = 15, k = 3, p = 1, s = 1$)	$224 \times 224 \times 3$	$224 \times 224 \times 15$
Conv2d($f = 30, k = 3, p = 1, s = 1$)	$224 \times 224 \times 15$	$224 \times 224 \times 30$
Maxpool($k = 2, p = 1, s = 1$)	$224 \times 224 \times 30$	$112 \times 112 \times 30$
Conv2d($f = 45, k = 3, p = 1, s = 1$)	$112 \times 112 \times 30$	$112 \times 112 \times 45$
Conv2d($f = 60, k = 3, p = 1, s = 1$)	$112 \times 112 \times 45$	$112 \times 112 \times 60$
Maxpool($k = 2, p = 1, s = 1$)	$112 \times 112 \times 60$	$56 \times 56 \times 60$
Conv2d($f = 75, k = 3, p = 1, s = 1$)	$56 \times 56 \times 60$	$56 \times 56 \times 75$
Conv2d($f = 90, k = 3, p = 1, s = 1$)	$56 \times 56 \times 60$	$56 \times 56 \times 90$
Maxpool($k = 2, p = 1, s = 1$)	$56 \times 56 \times 90$	$28 \times 28 \times 90$

Table 6. Layers corresponding to the type B CNN - Proposed architecture.

Layer	Input size	Output size
Conv2d($f = 105, k = 5, p = 1, s = 1$)	$28 \times 28 \times 180$	$28 \times 28 \times 105$
Maxpool($k = 2, p = 1, s = 1$)	$28 \times 28 \times 105$	$14 \times 14 \times 105$
Conv2d($f = 90, k = 5, p = 1, s = 1$)	$14 \times 14 \times 105$	$14 \times 14 \times 90$
Maxpool($k = 2, p = 1, s = 1$)	$14 \times 14 \times 90$	$7 \times 7 \times 90$
Conv2d($f = 75, k = 5, p = 1, s = 1$)	$7 \times 7 \times 90$	$7 \times 7 \times 75$
Maxpool($k = 2, p = 1, s = 1$)	$7 \times 7 \times 75$	$4 \times 4 \times 75$
Flatten	$4 \times 4 \times 75$	1200
Linear	1200	1000
Linear	1000	7

2. Testing different activation functions

The activation functions tested were the Rectified Linear Unit (ReLU), Leaky Rectified Linear Unit (LReLU), and Exponential Linear Unit (ELU). ReLU, as defined in Eq. 1, helps reduce the vanishing gradient problem and promotes sparse coding. However, it can cause issues when a large negative bias is learned, leading to constant zero outputs. To address this, Leaky ReLU, Eq. 2, allows a small gradient for negative inputs ($x < 0$). An alpha of 0.1 was chosen instead of the default 0.01 due to better accuracy in trials. Finally, ELU, as presented in Eq. 3, pushes mean unit activations closer to zero, which accelerates the model's training. An alpha of 1 was used, consistent with

Table 7. Designed units for the proposed CNN architecture.

(a) Type I	(b) Type II
Convolutional Layer(K=3,P=1,S=1)	Convolutional Layer (K=3,P=1,S=1)
Batch Normalization	Batch Normalization
Activation Function (ReLu,LReLu,ELU)	Activation Function (ReLu,LReLu,ELU)
Convolutional Layer(K=3,P=1,S=1)	Pooling Layer (K=2,P=1,S=1)
Batch Normalization	Dropout Layer (0.1)
Activation Function (ReLu,LReLu,ELU)	
Pooling Layer (K=2,P=1,S=1)	

previous studies that reported satisfactory results using this default value.

$$\text{ReLU}(x) = \max(0, x) \quad (1)$$

$$\text{LReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x < 0 \end{cases} \quad (2)$$

$$\text{ELU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x < 0 \end{cases} \quad (3)$$

3. Transfer learning using pre-trained CNN models

The state of the art in pre-trained CNN models is vast, posing a challenge due to the high complexity and longer training times of modern architectures. This study focused on three models-ResNet, DenseNet, and EfficientNet-to enhance peach leaf classification. To mitigate hardware and resource limitations, simpler variants of these models were chosen, reducing training times while still yielding satisfactory results, though more complex variants remain unexplored. For ResNet, the 50-layer variant was used, notable for its residual connections that map each layer’s input to its output. DenseNet, with 121 layers, improves performance by connecting each layer’s input to all previous layers, enhancing the flow of information and gradients. Lastly, EfficientNet B0 was tested, leveraging depthwise separable convolutions and lightweight blocks for high accuracy with reduced computational cost.

The pre-trained model pipeline is illustrated in Fig. 3. Unlike the pipeline in Fig. 2, no CNN block is added after the concatenation of feature maps due to the architectures of the pre-trained models. In this case, the output feature maps from each pre-trained model can no longer be reduced due to the network architectures. For this reason, no convolutional neural network block is included after the concatenation block. Instead, the concatenated feature maps are flattened and fed into a fully connected network consisting of two layers, with 1000 neurons in the input layer and 7 neurons in the output. The weights of these models were pre-trained on the ImageNet dataset. Data and source code are available at <https://github.com/Mateo-Coello/Siamese-Convolutional-Neural-Net> work.

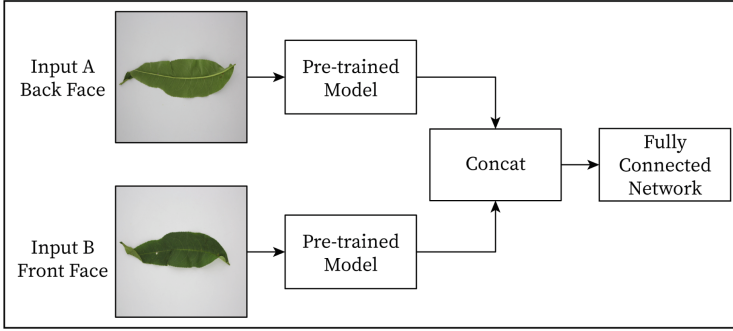


Fig. 3. Pipeline that considers the use of pre-trained models

4 Results and Discussion

4.1 Results of the Benchmark Architecture

The results of the five iterations considering the validation set for the architecture described in the benchmark paper are presented in Table 8. As mentioned previously, the benchmark architecture uses the ReLU activation function to initialize the weights of the convolutional layers. It is observed from the results of each iteration that the highest accuracy and lowest loss correspond to the fourth iteration, with a value of 0.9126 and 0.3637 respectively. Considering only the highest accuracy values obtained at the end of each iteration, the lowest one corresponds to the third iteration with a value of 0.8650 coupled with a loss of 0.8025. It is worth mentioning, that across all five iterations the highest accuracy obtained by the authors of the benchmark architecture, 0.9206, was not reached.

Table 8. Results of the benchmark architecture across the five iterations. HA: Highest Accuracy, A: Accuracy, LL: Lowest Lost, L: Lost.

Iteration	Results
1	HA: 0.8880, L: 0.4664
	LL: 0.4553, A: 0.8730
2	HA: 0.8730, L: 0.6199
	LL: 0.4418, A: 0.8730
3	HA: 0.8650, L: 0.8025
	LL: 0.7928, A: 0.8492
4	HA: 0.9126, LL: 0.3637
5	HA: 0.9048, L: 0.6012
	LL: 0.4366, A: 0.8968

4.2 Results of the Proposed Architecture

The results for the validation set of the proposed architecture were classified according to the activation function use and are presented in Table 9. For the ReLU activation function, it is observed that the highest accuracy achieved was 0.9444 with a loss of 0.3851 on the third iteration. The lowest loss of all iterations was also achieved on the third iteration, with a value of 0.3428 and an accuracy of 0.8809. The highest loss belongs to the first iteration with a value of 0.4827. Continuing, the LReLU activation function shows its highest accuracy with a value of 0.9444 and a loss of 0.3851 at the third iteration. Likewise, the lowest loss, 0.2882, with an accuracy of 0.9286 was also reached on the third iteration. Finally, the ELU activation function’s highest accuracy was 0.9286 with a 0.4201 loss achieved on the fourth iteration. In this same iteration, the lowest loss reached a value of 0.2723 and an accuracy of 0.9047.

Table 9. Results of the proposed model for the three activation functions. HA: Highest Accuracy, A: Accuracy, LL: Lowest Lost, L: Lost.

Iteration	Results		
	ReLU	LReLU	ELU
1	HA: 0.9048, L: 0.4827	HA: 0.9286, L: 0.3978	HA: 0.9286, L: 0.4201
	LL: 0.4729, A: 0.9048	LL: 0.3967, A: 0.8809	LL: 0.2723, A: 0.9047
2	HA: 0.9048, L: 0.3405	HA: 0.8730, L: 0.5125	HA: 0.9126, L: 0.5688
	LL: 0.3156, A: 0.9048	LL: 0.5097, A: 0.8571	LL: 0.4500, A: 0.8016
3	HA: 0.9444, L: 0.3851	HA: 0.9444, L: 0.3161	HA: 0.8809, L: 0.7343
	LL: 0.3428, A: 0.8809	LL: 0.2826, A: 0.9286	LL: 0.4837, A: 0.8254
4	HA: 0.9365, LL: 0.2882	HA: 0.9126, LL: 0.3670	HA: 0.9206, L: 0.4575
	–	–	LL: 0.3692, A: 0.9127
5	HA: 0.9126, L: 0.4457	HA: 0.9286, LL: 0.4529	HA: 0.9047, LL: 0.6045
	LL: 0.3465, A: 0.8730	–	–

4.3 Accuracy Comparison Between the Benchmark and Proposed Model

To compare how the proposed architecture improves the accuracy over the validation set regarding the benchmark model, its ReLU variant was considered. Again, it is necessary to emphasize that all iterations are independent, meaning the model is trained from zero at each iteration which does not affect if their order is changed. In this way, the highest accuracy across all iterations of both architectures was plotted in ascending order as shown in Fig. 4. The accuracy improvement across all iterations shows that indeed the proposed architecture in its ReLU variant has outperformed the benchmark model. Moreover, if the mean

of the highest accuracy values of both architectures is considered (benchmark mean: 0.8888, proposed ReLU variant mean: 0.9206), then is observed that the proposed architecture in its ReLU variant shows an accuracy improvement of 3.67%.

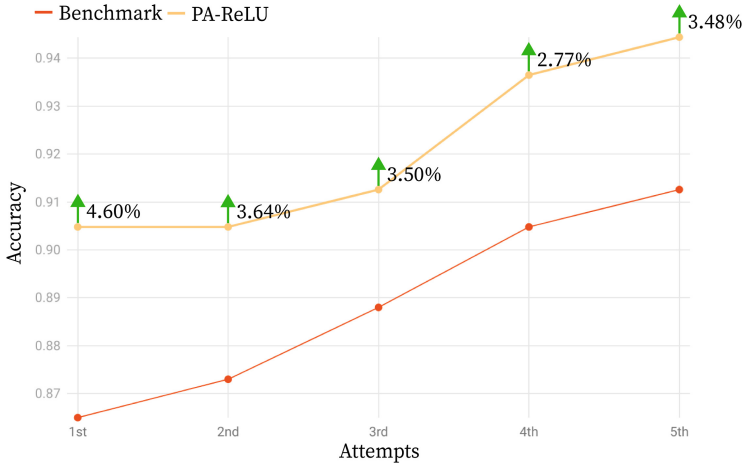


Fig. 4. Accuracy gain of proposed ReLU variant over benchmark architecture

4.4 Accuracy Comparison Between Proposed Architecture Variants

The maximum accuracy of the validation set across the five iterations for the three variants of the proposed model is illustrated in Fig. 5. This graph presents only the highest accuracy values across all five iterations. In the first iteration, the accuracy improvement for the LReLU variant is 3.64% for ReLU and 0.90% for ELU. Continuing, it is noticed that the LReLU variant has improved abruptly in the second iteration, showing an accuracy improvement of 0.86% in contrast to the other variants and an overall increase of 4.64% to the previous iteration. In the third iteration, LReLU still shows a higher accuracy over the other variants with an improvement of 1.75%.

However, at iteration number four, the ReLU variant shows the highest accuracy compared to the other variants. In this iteration, the accuracy improvement over ELU (lowest accuracy) is 1.72% for ReLU and 0.87% for LReLU. Lastly, at the fifth iteration both the ReLU and LReLU variants reached the highest accuracy across all five iterations, with a value of 0.9444 and an accuracy improvement of 1.70% compared to the ELU variant's highest accuracy of 0.9286. In this sense, no model demonstrates to outperform the others by a significant margin.

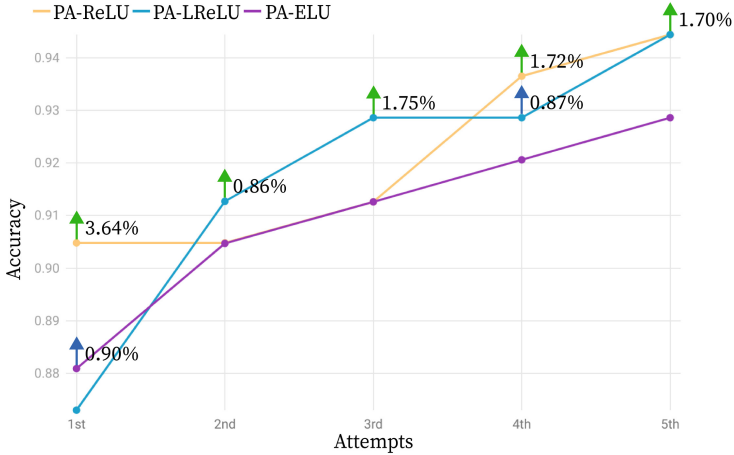


Fig. 5. Accuracy comparison between proposed architecture variants

4.5 Results Obtained with Each Pre-trained Model

The results for all pre-trained models are presented in Table 10. Starting with DenseNet121 model results, across all five iterations, the highest accuracy is 0.9603 obtained on the first iteration with a loss of 0.13977. In the following iterations and considering only the highest accuracy values, it is noticed that the minimum accuracy of all is 0.9365 reached in the fifth iteration. The lowest loss across corresponds to the fourth iteration with a value of 0.1167 and is accompanied by an accuracy of 0.9524. Moreover, regardless of the iteration, it is noticed that the final loss is always less than 0.2.

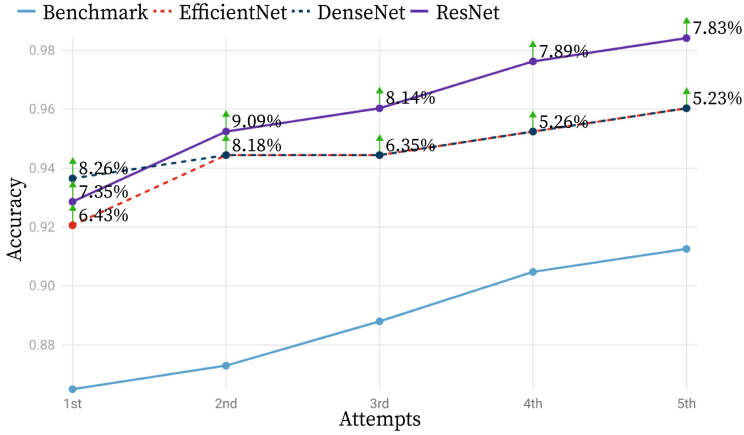
Among all the experiments presented in this article, ResNet50 obtained the highest accuracy and lowest loss results, with a value of 0.9841 and 0.0576 respectively. Additionally, it also presents the second-highest accuracy with a value of 0.9762 coupled with a loss of 0.1177 obtained on the fifth iteration. The EfficientNetB0 model has the highest accuracy on the second iteration with a value of 0.9603 although its loss is very high at 0.6794. A rather better version of this model will be obtained in the third iteration with an accuracy of 0.9525 and a loss of 0.2148. Moreover, is observed that the EfficientNet model presents higher losses than the previous pre-trained models.

4.6 Accuracy Comparison Between All Models

From the above results, it is noticeable that the pre-trained models have achieved higher accuracy values than both the reference architecture and the proposed architecture. Focusing our attention on Fig. 6, the accuracy gain of the pre-trained models to the benchmark architecture is observed. The plot contains the highest accuracy values reached by the benchmark and pre-trained models in each of its iterations arranged in ascending order. It is observed that

Table 10. Results of the proposed model for the three activation functions. HA: Highest Accuracy, A: Accuracy, LL: Lowest Lost, L: Lost.

Iteration	Results		
	DenseNet	ResNet	EfficientNet
1	HA: 0.9603, L: 0.13977	HA: 0.9841, LL: 0.0576	HA: 0.9206, LL: 0.5908
	LL: 0.1166, A: 0.9524	–	LL: 0.5599, A: 0.8492
2	HA: 0.9444, LL: 0.1680	HA: 0.9286, LL: 0.3261	HA: 0.9603, LL: 0.6794
	–	–	LL: 0.3742, A: 0.9524
3	HA: 0.9444, L: 0.1866	HA: 0.9603, L: 0.1639	HA: 0.9524, L: 0.2148
	LL: 0.1815, A: 0.9286	LL: 0.1158, A: 0.9524	LL: 0.2126, A: 0.9206
4	HA: 0.9524, LL: 0.1167	HA: 0.9524, L: 0.2473	HA: 0.9444, LL: 0.1767
	–	LL: 0.1289, A: 0.9444	–
5	HA: 0.9365, LL: 0.1613	HA: 0.9762, LL: 0.1177	HA: 0.9444, LL: 0.2202
	LL: 0.1472, A: 0.9365	–	–

**Fig. 6.** Accuracy gain of pre-trained models over benchmark architecture

DenseNet121 and EfficientNetB0's highest accuracy values are similar except for the first one. As mentioned earlier ResNet50 model's accuracy outperforms the other models with the mean of its highest accuracy values being 0.9603. In contrast, DenseNet121 presents an accuracy mean of 0.9476 while EfficientNetB0's accuracy mean is 0.9444. As such, ResNet50 shows an increase of 8.04%, DenseNet121 outputs an improvement of 6.62%, and EfficientNetB0 has a gain of 6.26%.

Since the variants of the proposed architecture do not outperform the others, only the ReLU variant was compared with the pre-trained models. In this regard, Fig. 7 presents the highest accuracy values obtained at the end of each

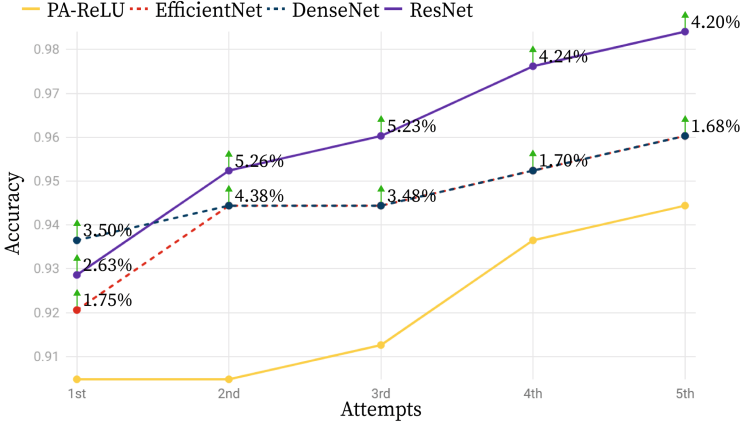


Fig. 7. Accuracy gain of pre-trained models over proposed ReLU variant architecture

iteration that follow the same rearrangement as in the previous plots, an ascending order. Overall, the accuracy gain for the proposed ReLU variant model is less significant than with the benchmark model, yet all the pre-trained models outperform it. By considering the mean of the highest accuracy values of the pre-trained models and proposed ReLU variant architecture, it is noticed that ResNet50 presents an improvement of 4.31%, DenseNet121 shows an increase of 2.93%, and EfficientNetB0 a gain of 2.52%.

5 Conclusions

In summary, this paper introduces an improved Siamese CNN that implements a custom architecture and pre-trained models to identify six varieties of peach leaves. In short, the proposed ReLU variant model compared to the benchmark architecture outperforms it, achieving a 3.67% accuracy improvement on the validation set. When comparing the different variants of the proposed architecture it was observed that no variant outperforms the others. Among the pre-trained models, ResNet50 consistently showed the highest accuracy gain over the benchmark architecture across all iterations, while DenseNet121 and EfficientNetB0 fell slightly behind. However, the accuracy reached by all pre-trained models outperforms both the proposed architecture variants and the benchmark model. Overall, this paper provides insights into the effectiveness of the proposed architecture and pre-trained models in identifying the front and back faces of peach leaves, which can be later applied for plant disease diagnosis and monitoring.

6 Potential Applications

Agriculture represents an important percentage of Ecuador's gross domestic product (GDP). According to statistics of the Central Bank of Ecuador [5],

the agriculture sector represented approximately 8% of the total GDP in 2022, with a total value of \$8,713,905. The distribution of this percentage is subdivided/comprised of several categories. A tool such as TradeMap highlights these categories generating statistics based on the information provided by the Central Bank of Ecuador and the United Nations Commodity Trade database. The link provided in Appendix C shows the total value of exports of different categories of products. Focusing our interest in the category *Edible fruit and nuts; peel of citrus fruit or melons*, its percentage of contribution to the total amount of exports in 2022 was around 10.76% or \$3,807,889. This raises the question of which products are comprised within this category. Considering the same year, bananas contributed 9.96% to the 10.76% or what it is \$3,523,427. The second category contains strawberries, raspberries, blackberries, gooseberries, etc., which contributed 0.35% or \$123,055. The third most valued category consists of dates, figs, pineapples, avocados, guavas, mangoes, and mangosteens, which shared 0.31% or \$110,857.

Albeit the exports of fresh apricots or peaches share almost none or zero percent of the total exports, these are products that are well rooted in the local economy of Ecuador. The provinces where peach production is concentrated are Azuay, Cotopaxi, Chimborazo, Imbabura, Pichincha, and Tungurahua. This is due to the conditions to properly cultivate the peach fruit, where an altitude between 2000 and 2800 masl is recommended, which is characteristic of these provinces. Several news articles comment that only one hectare of peach crops can produce tens of tons of such product. On the other hand, depending on the export category the restrictions imposed by each country vary, where pest control is essential for agricultural products like fruits and vegetables. As a result, a potential application of this work would be to further the classification capabilities of the CNN models by considering datasets where pests concentrate/spread in leaves.

Focusing such an application on Ecuador's agriculture requires the creation of a dataset of peach leaf varieties native to the country. This should include images properly tagged with the corresponding variety name and the pest that affects it. Embarking on such a study will require a proper knowledge of fruit agriculture and artificial intelligence to identify its scope, implement the corresponding models, and analyze if the results of the models are satisfactory enough in recognizing the pests. Likewise, the development of this application will determine the level of generalization and real application to other crops, as conceptually it is possible. Still, studies on this topic are lacking in Ecuador. Going one step further and applying this study to the fruit itself rather than its leaves will represent an interesting challenge as will also require a properly tagged dataset. Therefore, the applications are vast but studies should be carried out considering the economics (Ecuador's exports and local economy), computational resources (hardware to properly experiment), and the results to determine if a profitable service could be created.

Appendix

A Dataset Link

<https://drive.google.com/drive/folders/1rWCr9DrknoK0HKFhNRavCVgZ5UKjU3hi>.

B Implementation of Architectures Repository

<https://github.com/Mateo-Coello/Siamese-Convolutional-Neural-Network>.

C TradeMap Statistics Regarding Ecuador's Exports

https://www.trademap.org/Product_SelCountry_TS.aspx?nvpm=1|218||||TOTAL|||2|1|1|2|2|1|1|1|1|1.

References

1. Agarwal, M., Singh, A., Arjaria, S., Sinha, A., Gupta, S.: Toled: tomato leaf disease detection using convolution neural network. *Procedia Comput. Sci.* **167**, 293–301 (2020)
2. Akbar, M., Ullah, M., Shah, B., Khan, R.U., Hussain, T., Ali, F., Alenezi, F., Syed, I., Kwak, K.S.: An effective deep learning approach for the classification of bacteriosis in peach leave. *Front. Plant Sci.* **13**, 4723 (2022)
3. Ayala Niño, D., Cervantes Canales, J., García Lamont, F., Ayala de la Vega, J., Calderón Zavala, G.: A hybrid convolutional neural network for complex leaves identification. In: *Intelligent Computing Theories and Application: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part I* 16, pp. 291–301. Springer (2020)
4. Baganz, G.F., et al.: The aquaponic principle—it is all about coupling. *Rev. Aquac.* **14**(1), 252–264 (2022)
5. Banco Central del Ecuador: PIB nominal en millones de USD, anual 2022 (2022), <https://sintesis.bce.fin.ec/BOE/OpenDocument/2303281959/OpenDocument/opendoc/openDocument.jsp?logonSuccessful=true&shareId=0>. Accessed 21 Nov 2023
6. Beacham, A.M., Vickers, L.H., Monaghan, J.M.: Vertical farming: a summary of approaches to growing skywards. *J. Hortic. Sci. Biotechnol.* **94**(3), 277–283 (2019)
7. Camacho, J.C., Morocho-Cayamcela, M.E.: Mask R-CNN and YOLOv8 comparison to perform tomato maturity recognition task. In: Maldonado-Mahauad, J., Herrera-Tapia, J., Zambrano-Martínez, J.L., Berrezueta, S. (eds.) *Information and Communication Technologies*, pp. 382–396. Springer, Cham (2023)
8. Chen, P.C., Chiang, Y.C., Weng, P.Y.: Imaging using unmanned aerial vehicles for agriculture land use classification. *Agriculture* **10**(9), 416 (2020)
9. Espejo-García, B., Mylonas, N., Athanasakos, L., Fountas, S.: Improving weeds identification with a repository of agricultural pre-trained deep neural networks. *Comput. Electron. Agric.* **175**, 105593 (2020)

10. Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Vali, E., Fountas, S.: Combining generative adversarial networks and agricultural transfer learning for weeds identification. *Biosyst. Eng.* **204**, 79–89 (2021)
11. Figueroa, S., Morocho-Cayamcela, M.E., Pineda, I.: Identifying defective fruits and vegetables with hyper-spectral images: a brief tutorial. In: 2022 Third International Conference on Information Systems and Software Technologies (ICI2ST), pp. 21–27 (2022). <https://doi.org/10.1109/ICI2ST57350.2022.00011>
12. Hänni, N., Roy, P., Isler, V.: Apple counting using convolutional neural networks. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2559–2565. IEEE (2018)
13. Haupt, J., Kahl, S., Kowerko, D., Eibl, M.: Large-scale plant classification using deep convolutional neural networks. *CLEF (Working Notes)* **3** (2018)
14. Ivanová, H., Kalocaiová, M., Bolvansky, M.: Shot-hole disease on *Prunus persica*—the morphology and biology of *Stigmina carpophila*. *Folia Oecologica* **39**(1), 21 (2012)
15. Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y.: 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sens.* **10**(1), 75 (2018)
16. Kamilaris, A., Prenafeta-Boldú, F.X.: A review of the use of convolutional neural networks in agriculture. *J. Agric. Sci.* **156**(3), 312–322 (2018)
17. Kangune, K., Kulkarni, V., Kosamkar, P.: Grapes ripeness estimation using convolutional neural network and support vector machine. In: 2019 Global Conference for Advancement in Technology (GCAT), pp. 1–5. IEEE (2019)
18. Karypidis, E., Mouslech, S.G., Skoulariki, K., Gazis, A.: Comparison analysis of traditional machine learning and deep learning techniques for data and image classification. arXiv preprint [arXiv:2204.05983](https://arxiv.org/abs/2204.05983) (2022)
19. Khaki, S., Pham, H., Wang, L.: Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Sci. Rep.* **11**(1), 11132 (2021)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
21. Liu, H., Cocea, M., Ding, W.: Decision tree learning based feature evaluation and selection for image classification. In: 2017 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, pp. 569–574. IEEE (2017)
22. Lucero, V., Nobao, S., Morocho-Cayamcela, M.E.: Plant disease classification and severity estimation: a comparative study of multitask convolutional neural networks and first order optimizers. In: Lossio-Ventura, J.A., Valverde-Rebaza, J., Díaz, E., Muñante, D., Gavidia-Calderon, C., Valejo, A.D.B., Alatrística-Salas, H. (eds.) *Information Management and Big Data*, pp. 313–328. Springer, Cham (2022)
23. Organization, W.H., et al.: *The State of Food Security and Nutrition in the World 2020: Transforming Food Systems for Affordable Healthy Diets*, vol. 2020. Food & Agriculture Org. (2020)
24. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29**(9), 2352–2449 (2017)
25. Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., Homayouni, S.: Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **13**, 6308–6325 (2020)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)

27. Subramanian, M., Shanmugavadivel, K., Nandhini, P.: On fine-tuning deep learning models using transfer learning and hyper-parameters optimization for disease identification in maize leaves. *Neural Comput. Appl.* **34**(16), 13951–13968 (2022)
28. Takahashi, R., Matsubara, T., Uehara, K.: A novel weight-shared multi-stage CNN for scale robustness. *IEEE Trans. Circuits Syst. Video Technol.* **29**(4), 1090–1101 (2018)
29. UNDESA (United Nations Department of Economic and Social Affairs): *World Population Prospects: The 2017 Revision* (2017)
30. Virnodkar, S.S., Pachghare, V.K., Patil, V., Jha, S.K.: Remote sensing and machine learning for crop water stress determination in various crops: a critical review. *Precision Agric.* **21**(5), 1121–1155 (2020)
31. Zhong, G., Ling, X., Wang, L.N.: From shallow feature learning to deep learning: benefits from the width and depth of deep architectures. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* **9**(1), e1255 (2019)



Pancreas Segmentation Using SRGAN Combined with U-Net Neural Network

Mayra Elizabeth Tualombo¹ , Iván Reyes² , Paulina Vizcaíni-Imacaña² ,
and Manuel Eugenio Morocho-Cayamcela^{1,2}  

¹ School of Mathematical and Computational Sciences, Yachay Tech University, Hda. San José s/n y Proyecto Yachay, San Miguel de Urucuquí 10019, Ecuador
mayra.tualombo@yachaytech.edu.ec, mmorocho@yachaytech.edu.ec

² Universidad Internacional del Ecuador, Faculty of Technical Sciences, School of Computer Science, Quito 170411, Ecuador
{ireyes,pvizcaino,mamorochoca}@uide.edu.ec

Abstract. Pancreatic cancer, a serious disease, is challenging to detect early, hindering timely treatments. Medical imaging, especially computer tomography, is vital for early diagnosis. This study focused on accurate pancreatic segmentation in medical images using subtle changes in 2D U-Net architecture to aid in diagnosis and treatment. Additionally, we proposed using a super-resolution generative adversarial network (SRGAN) for better-resolution images. The proposed methodology considering the Dice-Sørensen coefficient (DSC), showed superior DSC values: 0.921163 ± 0.012 with subtle changes in U-Net architecture and 0.5354 ± 0.0002 DSC with SRGAN and the proposed U-Net considering low training time. Furthermore, the recall 0.6351 ± 0.0159 and precision 0.6422 ± 0.0056 were obtained. These results are important to the progression of medical imaging analysis for complex decision-making.

Keywords: Pancreas segmentation · Pancreatic tumour segmentation · Artificial intelligence · U-Net · SRGAN

1 Introduction

Pancreatic cancer is one of the most aggressive cancers and it is associated with 5% of cancer-related deaths [1]. According to the American Cancer Society, pancreatic cancer is the third leading cause of cancer-related death in the United States and corresponds with 466,003 cases worldwide [2,3]. Most patients diagnosed with pancreatic cancer did not show alarming symptoms until advanced metastasis pancreatic which led to less than 5 years of life expectancy in only 12.8% of cases during 2014–2020 [1]. One of the most problematic pancreas cancers is pancreatic ductal adenocarcinoma (PDAC) which accounts for more than 90% of pancreatic cancers [4].

Some factors, such as epidemiological trends, healthcare infrastructure, socioeconomic considerations, and genetic factors, are essential for understanding the progression of pancreatic cancer. Developed countries showed the highest

incident and mortality rates of pancreatic cancer [5]. However, South and Central America and Eastern Asia showed lower values [6]. Table 1 summarized those insights considering some countries.

Table 1. Pancreatic cancer deaths in some countries in 2020, according to American Cancer Society, WHO (World Health Organization), and Instituto Nacional de Estadística y Censos (INEC).

Country	Number of deaths in 2020
United States [7]	35,000
India [7]	32,812
Brazil [8]	11,999
China [7]	11,537
United Kingdom [7]	9000
Mexico [8]	4489
Ecuador [9]	562

When pancreatic cancer is detected, it often presents at an advanced stage, spreading to other parts of the body [4]. The early diagnosis of pancreatic cancer can boost the opportunity for long-term survival; however, patients are typically diagnosed at the right time incidentally which corresponds to less than 5% of the total cases creating health medical issues since more resources are needed [2].

The primary methods to detect pancreatic cancer are imaging and blood tests, followed by endoscopic procedures, genetic testing, and exploratory surgery. Imaging tests with radiographic images of pancreatic cancer are tedious and can lead to a mistaken diagnosis due to intra-operator ability [10]. Radiologists emphasize the need to enhance the visual analysis of pancreatic structure using tomography scans, as even subtle changes in size, shape, and structure may signal underlying pancreatic cancer issues [10]. Assessing the pancreas volume manually is prone to errors for each examiner. However, this challenge can be addressed by employing a computer-aided diagnosis system based on artificial intelligence (AI) [12]. Therefore, accurate segmentation of the human pancreas in medical imaging data is essential.

The techniques for detecting and segmenting pancreas and pancreatic tumors have been advancing, driven by the powerful capabilities of AI in accurately identifying objects within images [10]. Convolutional neural networks (CNNs) are used for segmenting, detecting, and localizing cancerous tumors helping with diagnosis and improving computer-assisted surgery techniques. The most widely recognized neural convolutional architecture is U-Net, which was designed for segmenting biomedical images.

Pancreatic cancer represents a challenge since (1) the shape of the pancreas is more variable and smaller than other types of organs considering that the pancreas is an approximately 15 cm long, spongy, and tube-shaped organ; (2)

the pancreas can move unpredictably within the retroperitoneal space affecting image processing; (3) poor contrast around boundaries of the pancreas in computed tomography (CT).

This article delved into the configuration of U-Net, a segmentation method designed to preserve CT image correlations and improve resolution. It also explored how subtle structural changes can enhance metrics such as the Dice-Sørensen coefficient (DSC). Furthermore, recognizing the significance of CT image resolution in the pancreatic medical field, super-resolution was introduced using a combined approach of the super-resolution (SR) generative adversarial networks (SRGANs) and U-Net.

2 Related Work

Image segmentation and resolution are important tasks in medical imaging. Pancreas segmentation from CT scans is one of the most challenging organ segmentation [11, 17, 44]. The U-Net architecture was developed to improve the efficiency of segmenting disease-affected regions. Krithika and Suganthi [13] adopted the baseline model of the U-net network and implemented a generalized pooling method to address the overfitting issue by adapting a pooling kernel's weights. Also, Zeng et al. [14] replaced the classical convolutional layers in the baseline method with RI-Blocks (Residual-Inception block) including residual and inception modules in the encoding phase. The experimental results in [14] showed a better performance than the classical U-Net. Many articles were based on the U-Net as a baseline model, making it more complex and sophisticated to produce better segmentation. However, sometimes complicating a model does not mean better metric evaluation results.

Some authors explained the need for a two-dimensional (2D) approach to effectively segment small areas of interest, taking into account the correlation between images. Li et al. [15], proposed an H-DenseUNet framework to explore and segment liver and tumors where the H-DenseUNet network combines 2D and three-dimensional (3D) networks, transforming the volumetric shape approach. The methodology ranked 1st on lesion segmentation in the 2017 LiTS Leaderboard [15].

On the other hand, resolution imaging was persuaded by Ledig et al. [16] who introduced the first GAN-based SR framework. Many of the commonly used SR techniques are non-machine learning-based methods. Sood et al. [18] proposed the implementation of SRGAN to produce SR versions of low resolution (LR)-MRI, considering that perceptual quality and edge fidelity are more important to a radiologist. Additionally, Tan et al. [19] combined the SRGAN architecture with meta-learning called Meta-SRGAN on the multimodal brain tumor segmentation and demonstrated a more efficient network of generating SR brain MRI images. Combining the segmentation approach with image resolution is important since well-segmented images with defined structures for visual inspection can help clinicians and radiologists.

3 Medical Image Analysis

Medical image analysis provides visual information about the human body. X-ray, CT, MRI, positron emission tomography (PET), and ultrasound are various imaging techniques utilized by radiologists and clinicians for precise diagnostic purposes [20]. CT and MRI are the two most common modalities for pancreas imaging [21].

3.1 Pancreas Segmentation with U-Net Network

Image segmentation is integrated with computer image processing to emphasize specific features that require highlighting. Segmentation involves dividing the image into interested correlated sub-regions which can be used to examine possible areas of interest to improve medical diagnosis [10]. The essential aspect of image segmentation involves transforming the image into an understandable format, enabling convenient utilization and analysis [22].

Organ segmentation specifically pancreas segmentation can be divided into two subgroups. Firstly, top-down multi-atlas registration and label fusion consider prior knowledge of the shape [23]. Secondly, bottom-up and deep learning methods show better results since CNNs avoid problems with blurred boundaries of the pancreas [24].

Deep learning methods also can be divided into two subgroups depending on the technique: (1) 2D CNNs and (2) 3D CNNs methods. U-Net network corresponds to the first group since the segmentation can be done even with a small input training dataset [25]. The U-Net network comprises two essential components: the down-sampling section, referred to as the extraction part, and the up-sampling component, referred to as the integration part [26]. Also, the division in the U-Net network can be understood as an encoder-decoder relation. On the other hand, in the second group, a 3D variant of the U-Net, known as the 3D U-Net, is employed.

3.2 CT Image Resolution Using SRGAN

Some deep neural networks such as GANs are used to obtain better image quality by removing noise [10]. This fact has captured the attention of medical experts since its applications in the biomedical field. Since the pancreas is attached to organs and occupies less than 2% of the entire CT volume, GANs-based methods have demonstrated significant improvement [27].

GANs are a class of AI algorithms used in unsupervised machine learning that consists of two competing networks, a discriminator and a generator [28]. The generator produces samples resembling the corresponding ground truths, while the discriminator endeavors to distinguish between the synthetic volumes and the original images, regardless of their degree of similarity [28].

High-resolution (HR) medical images can be obtained from LR medical images using SR methods [29]. Enhancing HR images from LR images is referred to as SR. This process aims to recover fine details and improve image quality

[30]. GANs combined with SR belong to a new classification named SRGANs where the generator is based on ResNet architecture that extracts features from LR images [29]. The initial segment of the network comprises multiple residual blocks, with each block containing two convolution layers of size 3×3 . Subsequently, a batch normalization (BN) layer is applied, followed by an activation function using the rectified linear unit (ReLU) function [16]. In the final part, two subpixel network modules are added to improve image resolution [16].

The discriminator follows the structure of the VGG-19 network, featuring eight convolutional layers [31]. The leaky rectified linear unit (LeakyReLU) function is applied as the activation function in the hidden layers [31]. Subsequently, the likelihood of whether the predicted image originates from the real HR image or the generated HR image is determined using a fully connected layer combined with a sigmoid activation function [31].

Some articles on applying GANs in medical imaging highlighted various GANs techniques. However, only a small group focused on the segmentation task due to its complexity. The structure of the remaining article was organized as follows. Firstly, an introduction to the dataset and image augmentation. Secondly, the description and configuration of U-Net and SRGAN methods used in the investigation. Then, the main results. Finally, a discussion considering the most relevant advances in the pancreas and pancreatic tumor segmentation, and a conclusion that summarizes the outstanding findings.

4 Methodology

4.1 Image Dataset

In this study, pancreatic CT scans from Medical Segmentation Decathlon (MSD) were used [32]. The dataset corresponds to 420 CT scans with the resolution of 512×512 pixels of portal-venous phase patients with pancreatic tumors labeled with background, pancreas, and tumor structures (Fig. 1). The original dataset is divided into 281 training and 139 testing sets.

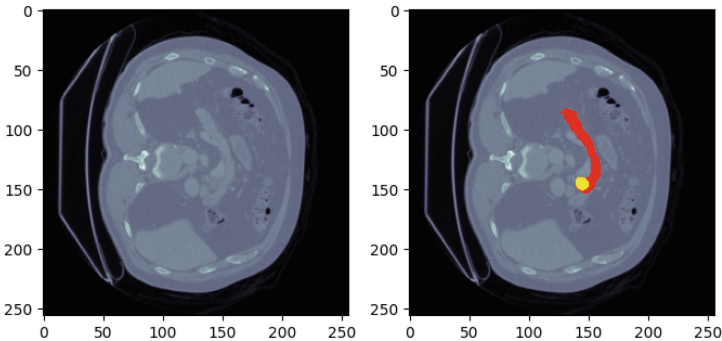


Fig. 1. Example of CT scan slice showing the pancreas in red and the tumor in yellow

This dataset includes three types of pancreatic tumors: intraductal mucinous neoplasms, pancreatic neuroendocrine tumors, and PDAC [23]. Each scan has a size of $n \times w \times h$ where n is the total number of slices, w is the horizontal dimension, and h is the vertical dimension.

Data division was performed based on the number of CT images used considering the execution time. Since U-Net perfectly works with a small dataset, as an example, for a subset of 40 CT images from the original training dataset, 2884 and 721 samples were considered for training and testing sets, respectively. The validation dataset was considered from the originally provided dataset of 139 CT images. Additionally, the preprocessed data was stored as 2D files to reduce the computational cost since reading a single slice is much faster than loading the complete NIFTI file. Finally, the images were normalized since CT images have a fixed range from -100 to 3071 .

Data Augmentation Class imbalance is a problem in semantic segmentation that can be solved using data augmentation. Some approaches such as resizing, rotation, and reflection are used in image data augmented setup to improve model prediction accuracy and reduce overfitting [33]. These methods are well-known as conventional augmentation methods. Although the original training set of 281 images was reduced for the training stage, during segmentation, the missing 241 images were added. Together with the data augmentation step, this creates a sufficiently reliable and diverse database.

4.2 Problem Definition

High-Resolution Images Since the generator of SRGANs gives pseudo-HR images depending on the training set, the discriminator of SRGANs tries to identify the error between the image created by the generator and the real ones [30]. This process can be summarized in the following equation:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{(\text{data})}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))], \quad (1)$$

where \mathbb{E} represents the expectation, G is the generator, D is the discriminator, x is the real sample, z is the sample of random noise, $p_{(\text{data})}(x)$ is the distribution of the real sample, $p_z(z)$ is the distribution of the noise, $D(x)$ is the output of the discriminator given x , and $D(G(z))$ is the output of the discriminator given a sample provided by the generator.

The loss function in (1) was defined considering the analysis developed by Liu et al. [30]. The construction considers:

1. The loss function \mathcal{L}_x is called the minimum absolute deviation:

$$\mathcal{L}_x = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I_{x,y}^{LR})^2). \quad (2)$$

Commonly, the mean square error (MSE) loss function is used in the construction of the SRGAN network to optimize the squared difference between each pixel of HR and LR images [34]. However, the weight of the MSE loss function to outliers is large. During the SR reconstruction process, even a small amount of noise in the image can significantly affect the reconstruction outcome. This is because many algorithms tend to amplify the noise during SR reconstruction [30]. Therefore, a regular term is needed in the optimization model.

2. Tvlss \mathcal{L}_{tv} is a regular term used in Eq. (5) which solves the difference between adjacent pixels in the image and is defined as follows:

$$\mathcal{L}_{tv} = \int_D \sqrt{u_x^2 + u_y^2} dx dy. \quad (3)$$

3. The adversarial loss component \mathcal{L}_{Gen}^{SR} is determined by how the discriminator assesses the output generated by the generator.

$$\mathcal{L}_{Gen}^{SR} = \sum_{n=1}^N -\log D_{\theta_D}(G_{\theta_D}(I^{LR})). \quad (4)$$

The overall loss function of SRGAN using Eqs. (2), (3), and (4), can be denoted as:

$$\mathcal{L}^{SR} = \mathcal{L}_x + 2 * 10^{-9} \mathcal{L}_{tv} + 10^{-3} \mathcal{L}_{Gen}^{SR}. \quad (5)$$

Pancreas Segmentation The pancreas segmentation problem is defined as follows:

Let,

$$X \in R^{n \times w \times h} \quad (6)$$

which is an arbitrary 3D scan data of a patient.

The segmentation model leads to a binary classification where $F(X) \in R^{n \times w \times h}$. The total pancreatic volume with its boundary will be denoted as Ω then,

$$F(X) = \begin{cases} 1, & \text{if } X \in \Omega, \\ 0, & \text{if } X \notin \Omega. \end{cases} \quad (7)$$

When the value in Eq. (7) corresponds to 1 the voxel belongs to Ω otherwise belongs to the background.

$F(X)$ corresponds to the ground truth and $\widehat{F}(x)$ corresponds to the outputs of the pixel-wise segmentation maps given by the three-dimensional scan. A mapping function ϕ can be constructed to minimize the difference between $F(X)$ and $\widehat{F}(X)$ or maximize the similarity.

One way to evaluate the similarity between two pixels is using the DSC. The DSC is calculated as the ratio of the intersection of segmented pixels compared to the sum of pixels in both segmentation and can be defined as follows:

$$\text{DSC (A, B)} = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (8)$$

where A represents the voxels obtained from $\widehat{F(X)}$ while B denotes the voxels in $F(X)$.

The value of DSC lies between 0 (with no superposition) and 1 (with superposition) and is easily converted into a percentage for understanding purposes. Finally, the use of equation (8) helps to define a loss function \mathcal{L} as follows:

$$\mathcal{L}(A, B) = 1 - \text{DSC}(A, B) \quad (9)$$

Additionally, recall measures the fractions of correctly predicted pixels in the total number of pancreas as follows:

$$\text{Recall}(A, B) = \frac{|A \cap B|}{|B|} \quad (10)$$

The equation for precision which describes the purity of the positive detections relative to the ground truth, can be defined as follows:

$$\text{Precision}(A, B) = \frac{|A \cap B|}{|A|} \quad (11)$$

4.3 Network Architecture

Model Considering the Specifications by Ronneberger et al. [35]: The down-sampling path comprises four 2×2 convolution (DoubleConv) layers with a ReLU activation function.

The ReLU layer sets any value to zero if the pixel value is less than zero. This layer is crucial because convolutional and BN layers frequently operate with non-uniform activation functions such as ReLU [33]. The ReLU function $g(x)$ can be defined as follows:

$$g(x) = \begin{cases} x, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \quad (12)$$

The DoubleConv is typically a combination of two convolutional (Conv2d) layers followed by activation functions, which help extract complex features from the input. In the Conv2d layer sliding convolution filters are applied to the input image. They convolve the received input by moving the filters both vertically and horizontally, while also computing the dot product of the weights and input [33]. Finally, a bias term is added.

A max-pooling layer with a stride of 2 was used to reduce the spatial dimension before every DoubleConv. Max-pooling layers do not learn directly, but they assist in reducing the number of features passed to subsequent connected layers and help minimize overfitting by reducing the number of parameters or weights [33]. The max-pooling technique identifies the biggest element in each pooling region [36]. The equation can be represented as follows:

$$f_{\max}(x) = \max(x_i x_j). \quad (13)$$

The m -th max-pooled band is composed of J -related filters,

$$pm = [p1, m, \dots, pj, m, \dots, pJ, m] \in R^J : \quad (14)$$

$$pj, m = \max(h_j, (m - 1)N + r). \quad (15)$$

where N ranges from 1 to R , representing a pooling shift that permits overlap between pooling regions when $N < R$ [36].

The fourth layer is the intermediate layer connecting the encoder and the decoder.

Up-sampling refers to increasing a feature map's spatial dimensions (w and h). In this case, the up-sampling path comprises three convolution layers (UpSampling2d). The UpSampling2d layer is a simple layer with no weights to double the dimensions of the input feature map [37]. A concatenation layer was added for skip connections followed by DoubleConv layers with ReLU activation function. Finally, the last layer produces the final segmentation map consisting of a Conv2d layer with a ReLU activation function [35]. The architecture has 7.8 M parameters in total and the structure is summarized in Fig. 2.

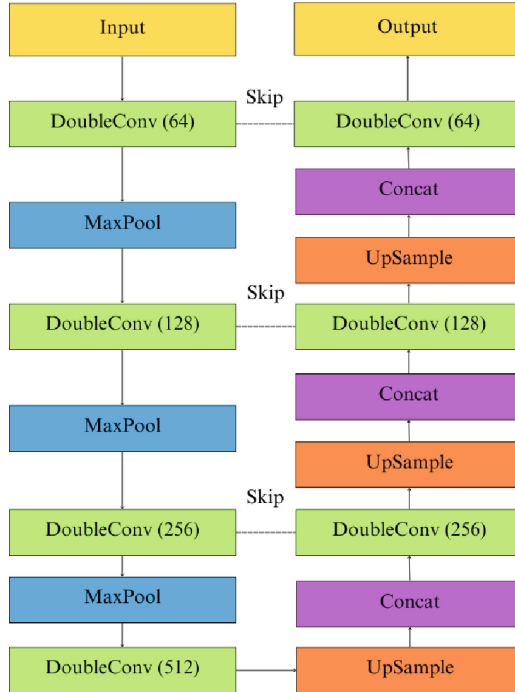


Fig. 2. Original structure of the U-Net model

Proposed Model Using Transposed Convolutional Layers: The class DoubleConv was actualized to be followed by a BN layer and ReLU activation function. BN was used to help the mean and variance of the parameters be the same, causing the input values to become stable getting a faster model during training [38].

The U-Net configuration was defined as follows: The encoder part constructs feature maps by applying convolutional layers. It iterated through the specified feature sizes, creating a DoubleConv block for each [33]. The decoder part reconstructed the segmented output, reversing the features' order to construct a series of transposed convolutional layers (ConvTrans) [33]. A ConvTrans layer is a convolution layer that up-samples feature maps and transposes the convolution [33]. Using this kind of layer the training time was reduced.

Considering the characteristics of the proposed problem, using Conv2d for feature detection was better since it shrinks the input shape. At the same time, ConvTrans performed better for feature extraction since it enlarges the input shape helping to upsample feature maps or generate an HR output [39, 40].

The bottleneck layer was implemented to reduce the feature size to the last specified feature size ensuring that the model captured relevant information. The final layer produced the output segmentation map. Consider the skip connections were stored during the encoder path and used during the decoder. The structure is summarized in Fig. 3.

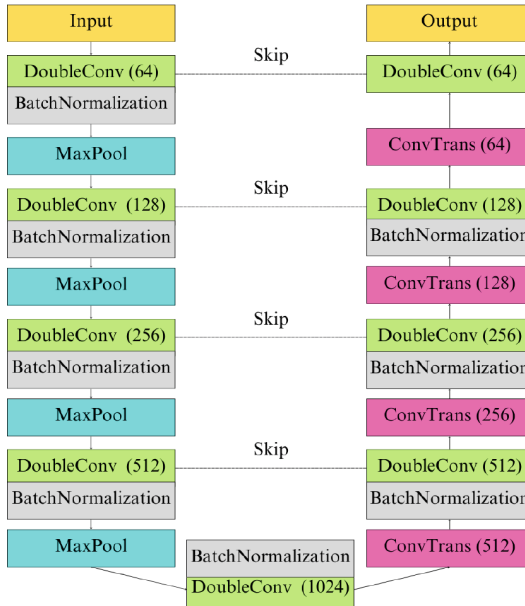


Fig. 3. Proposed model for U-Net using ConvTrans layer

Finally, the proposed U-Net model was incorporated into the SRGAN architecture, specifically within the generator component. U-Net was chosen for its ability to generate HR images. On the other hand, the discriminator was constructed using the Conv2d layer, followed by a Leaky ReLU. Additionally, there were six blocks consisting of Conv2d, BN, and Leaky ReLU, ultimately connected to a final Conv2d layer. The tumor segmentation process used the SRGAN model implemented with PyTorch Lightning. Finally, the architecture had 35.7 M parameters in total.

Figure 4 summarized the proposed methodology where data went to the data augmentation process using traditional augmentation methods. The resulting images were used to feed the proposed U-Net method which corresponded to the generator of the SRGAN architecture. Finally, those images were employed to train the discriminator getting the output to be evaluated using the metrics.

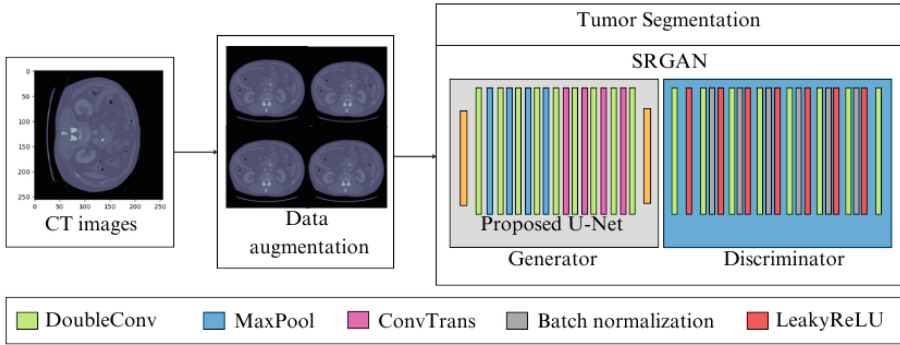


Fig. 4. The overall architecture of the proposed approach

5 Results

This study conducted a deep-learning analysis on an MSI Thin GF63 laptop equipped with a 12th-generation Core i7-12650H processor. The GPU used in this experiment is NVIDIA GeForce RTX. The networks were trained using the Adam optimizer (learning rate 0.0001) to minimize Eq. 9; batch size, 8; epoch, 5. Python (ver. 3.12.1, Python Software Foundation, USA) was used during the study. We conducted various sets of experiments to verify the availability of the proposed U-Net methodology. Then, we compared the model that combined SRGAN and U-Net with state-of-the-art pancreas segmentation methods.

Due to the specific characteristics of U-Net for segmentation tasks, various experiments were conducted starting from the basic model. Table 2 presents the evaluation results of the two segmentation models (U-Net configuration considering Ronneberger et al. [35] and the proposed architecture in Fig. 3).

Table 2. Evaluation metric for three pancreas segmentation models varying CT image number. Results are indicated as mean \pm standard deviation.

Method	DSC	Training time
U-Net model from [35] using 20 images	0.218874 ± 0.023	342 min
U-Net model from [35] using 40 images	0.589278 ± 0.016	798 min
U-Net model from [35] using 80 images	0.842328 ± 0.045	1523 min
Proposed U-Net model using 40 images	0.921 ± 0.012	304 min

Table 3. Compared our proposed model with state-of-the-art methods for pancreas segmentation.

Method	DSC	Recall	Precision
Attention U-Net and TAU-Net (hybrid model) are used for segmenting the pancreas and the tumor from CT images with auto-localization [23].	0.727	0.709 ± 0.21	0.701 ± 0.1
Pancreatic Segmentation via Ringed Residual Dense U-Net [41].	0.8832 ± 0.0284	Not available	Not available
The MAD-U-Net uses dense residual blocks and weighted binary cross-entropy to reduce the effects of intraclass inconsistency in pancreas segmentation [42].	0.8610 ± 0.0352	Not available	Not available
Application of a pyramid pooling U-Net model with integrated attention mechanism and inception mode in pancreatic tumor segmentation [43].	0.8790 ± 0.0419	0.9033 ± 0.1087	Not available
Pancreas segmentation with probabilistic map guided bi-directional recurrent U-Net (PBR-U-Net) [37].	0.85 ± 0.04	Not available	Not available
Automated pancreas segmentation and volumetry using deep neural network on computed tomography [12].	0.842	0.842	0.869
Improved version of the U-Net-based backbone network, AX-U-Net [45].	0.877 ± 0.038	0.909 ± 0.022	0.920 ± 0.061
Organ segmentation using two-dimensional U-Net [46]	0.57 ± 0.19	Not available	Not available
Organ segmentation using three-dimensional U-Net [46]	0.59 ± 0.15	Not available	Not available
SRGAN combined with proposed U-Net model	0.5354 ± 0.0002	0.6351 ± 0.0159	0.6422 ± 0.0056

We explored different image quantities since U-Net is known for its ability to work effectively with small amounts of data without compromising model performance. We observed that under controlled and equal conditions during training, the number of images truly enhanced the DSC. However, training time increased as well. On the other hand, the modifications made to the original model allowed us to achieve a better DSC value without compromising execution time. If we considered the factor of the number of images, in this case, 40, it

was evident that the modified model outperforms the original by 34%, while also reducing the execution time by half depicted in Table 2. This performance improvement can be attributed to the fact that the ConvTrans layer can also be understood as the gradient of the DoubleConv layer with respect to the input [33].

To evaluate the segmentation performance of the model developed in this article on the pancreas segmentation task, Table 3 summarized outstanding results in the state-of-the-art pancreas segmentation methods followed by the results of the proposed methodology. The DSC of 0.5354 indicated that the model performs moderately well in accurately segmenting areas of interest. The very narrow interval ± 0.0002 suggested that the model is consistently reliable in this metric across different samples. A value of 0.6351 for recall indicated that the model is quite good at identifying relevant areas. The narrow interval ± 0.0159 suggested low variability, indicating relatively stable recall performance. Finally, a value of 0.6422 for precision indicated that the model predicted a moderate amount of incorrect areas. The narrow interval ± 0.0056 suggested that precision is quite consistent across the samples.

The studies in Table 3 were designed to improve the deep learning model of automatic pancreas segmentation. They focused on three main aspects: (i) a comparison between two-dimensional and three-dimensional models; (ii) improvement of the network architecture; and (iii) enhanced image resolution.

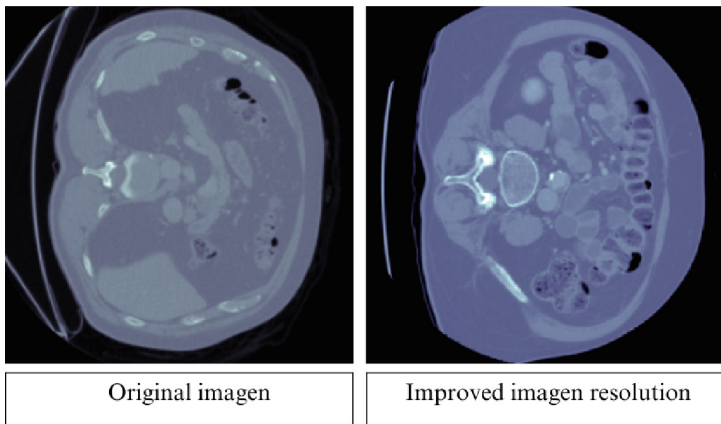


Fig. 5. Example of original and improved CT scan slice

Although the proposed model exhibits a lower value than the other studied methods depicted in Table 3, it is essential to consider that the training time varies significantly because the number of epochs is lower in our proposed methodology with a batch size of 8 and 60 epochs. This factor influenced the results since some authors consider a minimum training setting of over 500 epochs for deep learning NNs [12]. Figure 5 corresponded to the output from

the SRGAN combined with subtle changes in the U-Net structure. Despite the reduced number of epochs used in the experimentation, the results are favorable for enhancing the boundary definition of the anatomical structure, making it easier to identify.

6 Discussion

Deep learning methods have emerged as a promising approach, consistently outperforming conventional techniques in medical segmentation. NNs-based methods excel at identifying intricate patterns and features within datasets, enabling the capture of complex anatomical structures [10]. This study presented an automated deep-learning method for pancreatic and tumor segmentation using abdominal CT images from MSD.

In many biomedical applications, there is no need to use large image datasets for training a network [47]. Considering this, we only used a subset of 40 images. This decision was influenced by the fact that each image already contains repetitive structures with corresponding variations, given the complexity of pancreas anatomy. Additionally, the data augmentation process allows training the model with a small set of labeled data. This is useful since the open-access datasets about pancreatic tumors are limited.

The results obtained in Table 2 confirmed that the number of trainable parameters is directly proportional to the segmentation performance. This observation contrasted with findings by Lim et al. [12]. The experiments considered the U-Net model from Ronneberger et al. [35], used 7.8 M of parameters, while proposed changes used 31 M of parameters. Additionally, the ConvTrans layer played a vital role in time minimization in the training network. This characteristic was confirmed in the experimentation conducted by Paithane and Kakarwal [33] where the execution time was significantly reduced.

BN helped to reduce training time in numerous experiments. In our proposed model, using BN boosted the training time, helping metrics in U-net experimentation. On the other hand, Liu et al. [30] analyzed an SR algorithm based on an attention mechanism and the SRGANs network. The authors deleted the original BN layer to improve network performance. However, this change in our proposal can affect the performance of U-Net. Figure 5 showed the input and the output of the SRGAN process which can be improved, augmenting the number of epochs in the training process.

Understanding the intricacies of pancreatic cancer for accurate diagnosis and effective treatment, the results obtained in the current article are promising considering that U-Net showed better results in other types of cancer than pancreatic cancer. Zettler et al. [46] conducted experiments using 2D U-Net and 3D U-Net on organs such as the liver, right kidney, left kidney, spleen, and pancreas. While the 2D U-Net configuration yielded better results for all organs except the pancreas, the experiments with the 3D U-Net configuration were unfavorable for the pancreas, resulting in a DSC of 0.59 ± 0.15 . These results were attributed to the limited availability of axial training slices for the pancreas, as its elongation is not prominent along the z-axis [46].

7 Conclusion

The article focused on the mathematics involved in parameter selection for segmentation using neural networks, such as U-Net. The results were promising, particularly given the advantage of using relatively small datasets, which is a common challenge in biomedical analysis. On the other hand, neural networks like SRGANs have been studied for various computer vision problems. However, their complexity and significant computational requirements make them increasingly less accessible.

Despite achieving favorable results for the proposed U-Net model, it is possible that incorporating changes such as dropout layers would have led to even better outcomes by preventing overfitting. However, since favorable results were obtained, the experimentation concluded. Techniques like hyperparameter tuning could address uncertainties in parameter selection. Therefore, this article focused on the mathematical implementation, utilizing enhancements such as the gradient of the ConvTrans layer.

Since the images captured by the generator, which comes from the proposed U-Net structure, must correspond to HR images, it ensures better results for both the discriminator and the entire SRGAN. Using a generator pre-trained on an HR image database could yield more favorable outcomes. However, the limited availability of open data poses challenges during the experimentation process.

Our study has certain limitations that can be addressed in future research. First, increasing the training time could lead to better resolution when using the SRGAN methodology. Second, since the U-Net architecture showed promising results, it should be tested in combination with another neural network structure to further improve image resolution. Third, additional configurations, such as the dropout technique or the exponential linear unit (ELU) activation function, should be explored to enhance the effectiveness of the U-Net network. Finally, some authors cited in the bibliography argued that traditional data augmentation methods should be thoroughly studied to identify the optimal combination, as they can significantly impact model performance. We also recommend exploring mixup and patching techniques, as they could yield better results.

References

1. Zhao, Z., Liu, W.: Pancreatic cancer: a review of risk factors, diagnosis, and treatment. *Technol. Cancer Res. Treat.* **19**, 1533033820962117 (2020)
2. Halbrook, C.J., Lyssiotis, C.A., di Magliano, M.P., Maitra, A.: Pancreatic cancer: advances and challenges. *Cell* **186**(8), 1729–1754 (2023)
3. Partyka, O., et al.: Overview of pancreatic cancer epidemiology in Europe and recommendations for screening in high-risk populations. *Cancers* **15**(14), 3634 (2023)
4. Hu, J.X., Zhao, C.F., Chen, W.B., Liu, Q.C., Li, Q.W., Lin, Y.Y., Gao, F.: Pancreatic cancer: A review of epidemiology, trend, and risk factors. *World J. Gastroenterol.* **27**(27), 4298 (2021)










5. Ilic, M., Ilic, I.: Epidemiology of pancreatic cancer. *World J. Gastroenterol.* **22**(44), 9694 (2016)
6. Bosetti, C., Bertuccio, P., Negri, E., La Vecchia, C., Zeegers, M.P., Boffetta, P.: Pancreatic cancer: overview of descriptive epidemiology. *Mol. Carcinog.* **51**(1), 3–13 (2012)
7. American Cancer Society: Key Statistics for Pancreatic Cancer. American Cancer Society. <https://www.cancer.org/es/cancer/tipos/cancer-de-pancreas/acerca/estadisticas-clave.html> (2024)
8. World Health Organization: Cancer. World Health Organization. <https://www.who.int/es/news-room/fact-sheets/detail/cancer> (2022)
9. Instituto Nacional de Estadística y Censos: Search: Cancer. Instituto Nacional de Estadística y Censos. <https://www.ecuadorencifras.gob.ec/search/cancer/> (2020)
10. Ghorpade, H., et al.: Automatic segmentation of pancreas and pancreatic tumor: a review of a decade of research. *IEEE Access* (2023)
11. Macancela, C., Morocho-Cayamcela, M., Chang, O.: Deep reinforcement learning for efficient digital pap smear analysis. *Computation* **11** (2023), <https://www.mdpi.com/2079-3197/11/12/252>
12. Lim, S.H., Kim, Y.J., Park, Y.H., Kim, D., Kim, K.G., Lee, D.H.: Automated pancreas segmentation and volumetry using deep neural network on computed tomography. *Sci. Rep.* **12**(1), 4075 (2022)
13. AnbuDevi, K.A.M., Suganthi, K.: Review of semantic segmentation of medical images using modified architectures of UNET. *Diagnostics* **12**(12), 3064 (2022)
14. Zeng, Z., Xie, W., Zhang, Y., Lu, Y.: RIC-Unet: an improved neural network based on Unet for nuclei segmentation in histology images. *IEEE Access* **7**, 21420–21428 (2019)
15. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018)
16. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690 (2017)
17. Román, K., Llumiquinga, J., Chancay, S., Morocho-Cayamcela, M.: Hyperparameter tuning in a dual channel U-Net for medical image segmentation. *Inf. Commun. Technol.* 337–352 (2023)
18. Sood, R., Topiwala, B., Choutagunta, K., Sood, R., Rusu, M.: An application of generative adversarial networks for super resolution medical imaging. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 326–331. IEEE (2018)
19. Tan, C., Zhu, J., Lio', P.: Arbitrary scale super-resolution for brain MRI images. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 165–176. Springer, Cham (2020)
20. Sarvamangala, D.R., Kulkarni, R.V.: Convolutional neural networks in medical image understanding: a survey. *Evol. Intell.* **15**(1), 1–22 (2022)
21. Proietto Salanitri, F., Bellitto, G., Irmakci, I., Palazzo, S., Bagci, U., Spampinato, C.: Hierarchical 3d feature learning for pancreas segmentation. In: *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pp. 238–247. Springer (2021)
22. Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K.: Medical image analysis using convolutional neural networks: a review. *J. Med. Syst.* **42**, 1–13 (2018)

23. Mahmoudi, T., et al.: Segmentation of pancreatic ductal adenocarcinoma (PDAC) and surrounding vessels in CT images using deep convolutional neural networks and texture descriptors. *Sci. Rep.* **12**(1), 3092 (2022)
24. Li, Y., Liu, D., Li, H., Li, L., Wu, F., Zhang, H., Yang, H.: Convolutional neural network-based block up-sampling for intra frame coding. *IEEE Trans. Circuits Syst. Video Technol.* **28**(9), 2316–2330 (2017)
25. Deng, Y., Lan, L., You, L., Chen, K., Peng, L., Zhao, W., Zhou, X.: Automated CT pancreas segmentation for acute pancreatitis patients by combining a novel object detection approach and U-Net. *Biomed. Signal Process. Control* **81**, 104430 (2023)
26. Xiao, Z., Liu, B., Geng, L., Zhang, F., Liu, Y.: Segmentation of lung nodules using improved 3D-UNet neural network. *Symmetry* **12**(11), 1787 (2020)
27. Yan, Y., Zhang, D.: Multi-scale U-like network with attention mechanism for automatic pancreas segmentation. *PLoS ONE* **16**(5), e0252287 (2021)
28. Li, M., Lian, F., Guo, S.: Automatic pancreas segmentation using double adversarial networks with pyramidal pooling module. *IEEE Access* **9**, 140965–140974 (2021)
29. Ahmad, W., Ali, H., Shah, Z., Azmat, S.: A new generative adversarial network for medical images super resolution. *Sci. Rep.* **12**(1), 9533 (2022)
30. Liu, B., Chen, J.: A super resolution algorithm based on attention mechanism and SRGAN network. *IEEE Access* **9**, 139138–139145 (2021)
31. Xiong, Y., Guo, S., Chen, J., Deng, X., Sun, L., Zheng, X., Xu, W.: Improved SRGAN for remote sensing image super-resolution across locations and sensors. *Remote Sens.* **12**(8), 1263 (2020)
32. Medical Decathlon: Retrieved June 9, 2024, from <http://medicaldecathlon.com/> (n.d.)
33. Paithane, P.M., Kakarwal, S.N.: Automatic pancreas segmentation using a novel modified semantic deep learning bottom-up approach. *Int. J. Intell. Syst. Appl. Eng.* **10**(1), 98–104 (2022)
34. Kant, M., Chaurasia, S., Sharma, H.: Contribution analysis of scope of SRGAN in the medical field. In: *Data Engineering for Smart Systems: Proceedings of SSIC 2021*, pp. 341–352. Springer, Singapore (2022)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv./abs/1505.04597* (2015)
36. Zafar, A., Aamir, M., Mohd Nawi, N., Arshad, A., Riaz, S., Alruban, A., Dutta, A.K., Almotairi, S.: A comparison of pooling methods for convolutional neural networks. *Appl. Sci.* **12**(17), 8643 (2022)
37. Li, J., Lin, X., Che, H., Li, H., Qian, X.: Pancreas segmentation with probabilistic map guided bi-directional recurrent UNet. *Phys. Med. Biol.* **66**(11), 115010 (2021)
38. Santurkar, S., Tsipras, D., Ilyas, A., Madry, A.: How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, vol. 31 (2018)
39. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2528–2535. IEEE (2010)
40. Shi, W., Caballero, J., Theis, L., Huszar, F., Aitken, A., Ledig, C., Wang, Z.: Is the deconvolution layer the same as a convolutional layer? *arXiv preprint arXiv:1609.07009* (2016)
41. Lu, L., Jian, L., Luo, J., Xiao, B.: Pancreatic segmentation via ringed residual U-Net. *IEEE Access* **7**, 172871–172878 (2019)

42. Li, W., Qin, S., Li, F., Wang, L.: MAD-UNet: a deep U-shaped network combined with an attention mechanism for pancreas segmentation in CT images. *Med. Phys.* **48**(1), 329–341 (2021)
43. Zhang, Z., Tian, H., Xu, Z., Bian, Y., Wu, J.: Application of a pyramid pooling Unet model with integrated attention mechanism and Inception module in pancreatic tumor segmentation. *J. Appl. Clin. Med. Phys.* **24**(12), e14204 (2023)
44. Paucar, F., Bojorque, C., Reyes-Chacón, I., Vizcaino-Imacaña, P., Morocho-Cayamcela, M.: Towards accurate cervical cancer detection: leveraging two-stage CNNs for pap smear analysis. In: *Proceedings of the 19th International Conference on Software Technologies*, vol. 1: ICSOFT, pp. 219–227 (2024)
45. Yang, M., Zhang, Y., Chen, H., Wang, W., Ni, H., Chen, X., Li, Z., Mao, C.: AX-Unet: a deep learning framework for image segmentation to assist pancreatic tumor diagnosis. *Front. Oncol.* **12**, 894970 (2022)
46. Zettler, N., Mastmeyer, A.: Comparison of 2D vs. 3D U-Net organ segmentation in abdominal 3D CT images. arXiv preprint [arXiv:2107.04062](https://arxiv.org/abs/2107.04062) (2021)
47. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*, pp. 424–432. Springer (2016)



Using Artificial Intelligence and X-ray Images to Train and Predict COVID-19 and Pneumonia: Tool for Diagnosis and Treatment

Bryan Juárez-Gonzalez¹ , Fernando Villalba-Meneses¹ ,
Jonathan Cruz-Varela¹ , Andrés Tirado-Espín² ,
Paulina Vizcaíno-Imacaña³ , Carolina Cadena-Morejon² , Cesar Guevara⁴ ,
and Diego Almeida-Galárraga¹  

¹ School of Biological Sciences and Engineering, Universidad Yachay Tech, Urcuquí 100119, Ecuador

dalmeida@yachaytech.edu.ec

² School of Mathematical and Computational Sciences, Universidad Yachay Tech, San Miguel de Urcuquí 100119, Ecuador

³ Universidad Internacional del Ecuador, Quito, Ecuador

⁴ Centro de Investigación en Mecatrónica y Sistemas Interactivos MIST, Universidad Tecnológica Indoamérica, Machala y Sabanilla, Quito 170103, Ecuador

cesarguevara@uti.edu.ec

Abstract. In January 2022, Ecuador experienced a peak in COVID-19 cases, with 890,541 confirmed cases and 35,658 deaths. During 2019-2020, influenza and pneumonia also ranked among the top causes of death. Traditional chest X-rays and chest CT scans are commonly used for diagnosing COVID-19, but studies by Wong et al. indicate that chest X-rays are less sensitive compared to CT scans unless artificial intelligence is utilized. Tahir et al. highlighted that AI models such as U-net and neural networks achieve high diagnostic accuracy, with U-Net++ and ResNet18 models showing sensitivities above 99% and perfect specificity using a large dataset of 33,920 chest X-ray images. The rapid detection of symptoms could have helped in prioritizing critical care, potentially reducing deaths. In the current study, it is demonstrated that combining AI with chest X-rays can achieve a binary accuracy above 98% for COVID-19 detection using transfer learning with Xception, VGG16, and VGG19 neural networks. Using 27,052 chest X-ray images, the VGG19 model achieved an excellent F1-score of 98.53% for COVID-19 and normal class classification. The VGG19 model also performed well in multiclass classification with an F1-score above 89%. The study concludes that AI-enhanced diagnostic tools, such as VGG19, are valuable for hospitals in diagnosing COVID-19 and future improvements might include larger datasets and enhanced segmentation techniques.

Keywords: Artificial intelligence · COVID-19 · Pneumonia · chest X-ray · transfer learning · Xception · VGG16 · VGG19

1 Introduction

Ecuador presented a peak of COVID-19 cases in January 2022, with 890,541 confirmed cases and 35,658 deaths [18]. Also, influenza and pneumonia, not related to COVID-19, were among the top 5 causes of death in Ecuador during 2019, according to INEC Ecuador government statistics [12]. As consequence, we had lack of tools and resources in Ecuador and all over the world. Then, the need for a tool to predict whether or not a patient had COVID-19 appeared. Neural networks have proven to be very effective at learning complex image patterns. So, chest X-ray images that have complicated structures were used in this work. Our objective is to create and test state-of-the-art neural networks that have been used in the prediction of COVID-19 using chest X-rays and try to replicate and possibly improve the results that have been reported so far. Also, the models will be tested with different classes, and the results will be reported to conclude which model is better. The trained models could be used by Information Technology (IT) developers to create mobile apps and websites that can aid in the diagnosis of COVID-19.

1.1 Artificial Intelligence

Artificial intelligence (AI) is the most exciting and talked-about technology of the past ten years. According to John McCarthy, it is the process of making intelligent machines, especially by creating intelligent computer programs that mimic human intelligence [17]. AI is currently used in many industries, including business, healthcare, education, the military, and manufacturing [24].

In medical diagnosis, AI is being widely used. As reported in [23], searchers have used machine learning techniques to detect early Alzheimer's disease using MRI images and data. Surgical robots can be trained using AI to achieve precise control of medical equipment [23]. In breast cancer screening, AI can detect and classify breast lesions [7]. Using chest tomography scans, machine learning algorithms can predict the probability of developing lung cancer [4]. Real-time detection in combination with deep learning (DL) was used in the diagnosis of basal cell carcinoma [25]. Machine learning and DL are used in the detection of tuberculosis using chest X-ray images [10]. Also, DL is used to detect lung cancer, pneumonia, lung opacity, and COVID-19 [2]. Convolutional neural networks (CNNs), transfer learning, generative adversarial networks, recurrent neural networks, and attention mechanisms are techniques used in machine learning to detect patterns from images and extract information from data [16]. Convolutional neural networks are widely used in image processing and, recently, transfer learning which will be expanded further.

1.2 High Precision Models Using Chest X-rays

Neural networks are being used to accurately predict patterns in new images. Now, I will mention some of these frameworks that have had good results. For

example, the Bayesian deep neural network has an accuracy of 92.9% using chest X-ray images [9].

The DarkNet neural network was also used [19], obtaining an accuracy of 98.08% for binary classification (COVID vs. No-Findings) and 87.02% for multi-class (COVID vs. No-Findings vs. Pneumonia). They used 224 COVID-19 images, 700 for bacterial pneumonia, and 504 for normal patients. They used balancing methods to select only 500 images from bacterial pneumonia and 500 from normal. They also mentioned that the number of images that they used for COVID-19 was low due to the lack of positive cases at the time they conducted the study.

According to the study by [3], the authors used transfer learning models and obtained these results for accuracy in Table 1. We can observe that VGG-19 performed better than the other models for both binary and other classes.

Table 1. Several transfer learning model accuracies are reported. The binary classification used the COVID and Normal classes. The three-class classification used COVID, Normal, and Pneumonia

Model	Binary classification [%]	Three classes [%]
VGG-19	98.75	93.48
MobileNet	97.40	92.85
Inception	86.13	92.85
Xception	85.57	92.85
Inception ResNet v2	84.38	92.85

Sitaula et al. [21], used the VGG-16 transfer learning model with chest X-ray images. They used three datasets containing an average of 300 images for each class. The first dataset has 3 classes, the second has 4 classes, and the third has 5 classes. Classification accuracy values for datasets 1, 2, and 3 are reported in Table 2.

Table 2. VGG16 and VGG19 transfer learning model accuracies.

Model	Three classes [%]	Four classes [%]	Five classes [%]
VGG-16	79.58	85.43	87.49
VGG-19	74.84	82.83	85.00

Sitaula and its collaborators found that VGG-16 performed slightly better than VGG-19. It is important to mention that they used an internal testing dataset extracted from the training dataset. The correct method would be using an external testing dataset [13], which simulates better a real-life application with completely new images. This external dataset will ensure a more robust evaluation of the model and its ability to perform on unseen data.

2 Methods

The model was created using Jupyter Notebooks on the Google Colab and Kaggle platforms. Keras, tensorflow, pandas, numpy, sklearn, matplotlib, PIL, among other libraries were used. The dataset was selected taking into consideration the number of images and the resolution. The dataset was first checked for images that were not part of the Chest-Xray category and other unnecessary archives. Subsequently, the database was compressed and uploaded to a shared Google Drive folder. Then, the dataset was downloaded in each session of compiling by using the gdown library. So, the downloading only took 6 s per session using an average of 208 MB/s of download speed. Then the dataset was unzipped in the Colab and Kaggle platforms. Finally, the dataset was balanced to obtain the same amount of images for each class. The final size of the dataset is shown further. To train the model we stored the images of the dataset using Keras library. Then, we used grayscale images since color images did not show model accuracy improvement. Image augmentation, autotune, and randomizing were applied. A seed = 123 was selected to ensure reproducibility so that the randomization process would be done in a deterministic manner. In Fig. 1, we observe a summary of the work.

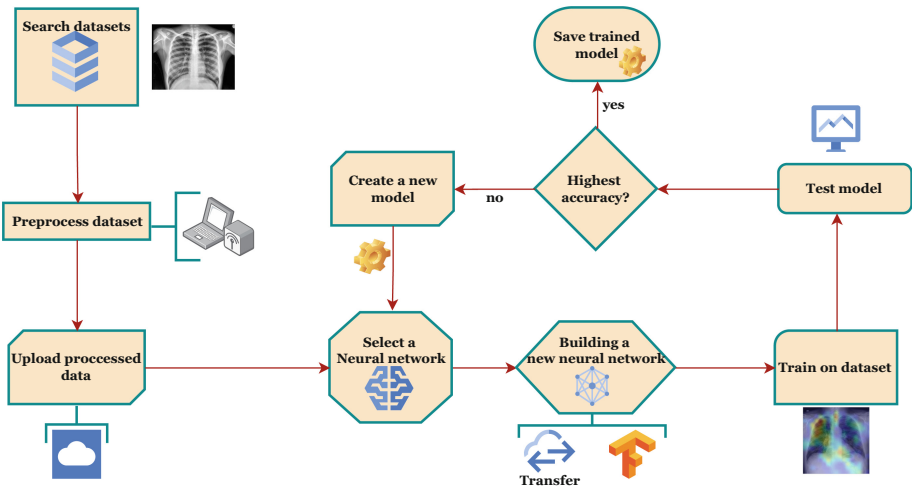


Fig. 1. Abstract of the process used to build and test the neural networks.

The public “Covid-19 radiography database” [5] was used in this experiment. The models used in this work were tested on binary and multiclass datasets that we created from the *Covid-19 radiography database* dataset. During the validation process, we used binary and categorical accuracy. Also, binary and categorical cross-entropy were used during the model compiling. The Adam optimizer was used. Heat maps and Grad-Cam were used to visualize the neural network activations. Receiver Operating Characteristic (ROC) was used to visually

determine the model performance. ROC is commonly used in binary classification using different threshold values to plot the data. In simpler terms, it's a way to visualize the performance of a binary classification model across different threshold settings. Threshold values determine the boundary in the prediction of classes.

2.1 Covid-19 Radiography Database

Kaggle provided the *Covid-19 radiography database* free dataset. The dataset contains a total of 27,052 images divided into 4 classes. The balancing of the dataset was used to improve the accuracy of the results. Also, data augmentation was applied to increase the number of images. Then, the dataset was divided into training, validation, and testing. A 70% was used for training, 15% for validations, and 15% for testing. Also, prefetching and shuffling were used to ensure the reproducibility of this work. The architecture of one of the neural networks is presented. Finally, the platforms we used and the hardware are shown. Table 3 contains the characteristics of the database [5]. In Fig. 2, a sample from the “COVID” class was extracted.

Table 3. Dataset characteristics. The number 3 in shape indicates that we have an RGB format for the initial images.

Characteristic	Value
Total of images	27,052
Normal class images	12,157
COVID class images	3815
Viral pneumonia images	5068
Lung opacity images	3815
Format	.jpg
Images initial size	400 × 400
Shape	(400,400,3)

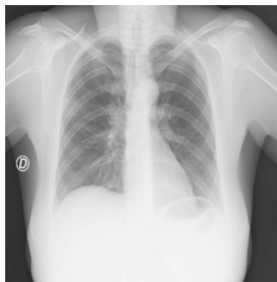


Fig. 2. Chest X-ray of a patient diagnosed positive for COVID-19 infection.

2.2 Balancing and Augmentation Techniques

The dataset was balanced to improve testing accuracy. The Python code counted the number of images in each folder class. Then the minimum number was 3815 images for the “COVID” folder. Next, the minimum number was used to randomly select samples from all the other classes. The process was done until all the classes had the same number of images as the minimal number. Then a new balanced dataset was created with a total of 15,260 images. Next, the balanced dataset was divided into training, validation, and testing. A 70% was used for training, 15% for validations, and 15% for testing. The percentages can be adjusted in the code as needed. Then, the images were loaded using Keras libraries. The dataset load parameters were: $height = 280$, $width = 280$, $batchsize = 32$, and $seed = 123$ “to ensure reproducibility when the batch size is randomized”. Color images were used in all the experiments due to the lack of significant changes when using grayscale samples.

In Fig. 3, augmentation technique is used using the Keras package with $random - flip = \text{“horizontal”}$ and $random - rotation = 0.1$. Then the models were constructed using the next structure of classes: COVID and Normal for binary classification, “COVID” “Normal” and “Viral Pneumonia”, and “COVID” “Normal” “Viral Pneumonia” and “Lung Opacity”.

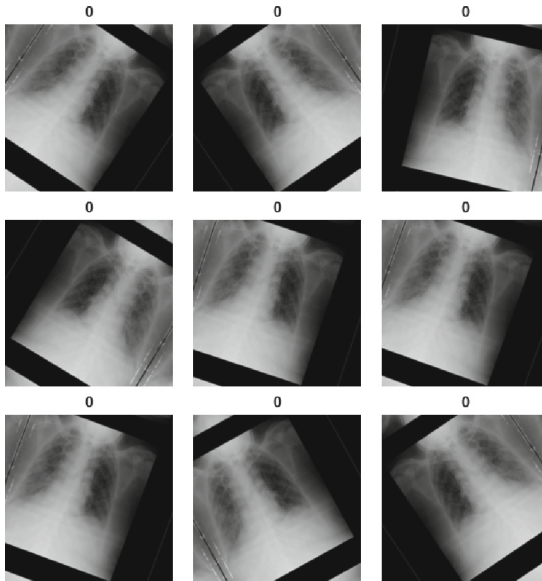


Fig. 3. Augmentation technique used on the dataset.

2.3 Software and Hardware Used in Our Experiment

Python code can be written and run in a collaborative and interactive environment with Google Colab, a free cloud-based platform. The capabilities of the

application, which are based on Jupyter Notebooks, include free GPU access for resource-intensive projects, real-time collaborative editing, easy sharing via Google Drive integration, pre-installed libraries for machine learning and data analysis, and support for exporting notebooks in multiple formats. Because of its convenience and accessibility, Colab is widely used in data science and machine learning applications. During the experiments, we used Google Colab, which used the Python 3 version. We also used the 2.14.0 “Tensorflow” version on the Jupyter Notebooks. We also used the Kaggle platform.

Google Cloud Platform (GCP) and Kaggle offer access to strong GPUs, such as T4 GPUs. Based on the NVIDIA Turing architecture, the T4 GPU is a good choice for workloads that require GPU acceleration, such as machine learning. In the free version, the one we used for this research, GCP offers 1 GPU with 16 GB of RAM of the type GDDR6. Kaggle offered up to 30 GB of RAM, doubling the GPU RAM size of Colab. Both the Colab and Kaggle platforms were used to run our code. The code we created for this work can be publicly accessed on Kaggle [14].

3 Experimental Results

Neural networks have proven to be very effective at learning complex image patterns. Our objective is to create and test state-of-the-art neural networks that have been used in the prediction of COVID-19 using chest X-rays and try to replicate and possibly improve the results that have been reported so far. Also, the models will be tested with different classes, and the results will be reported to conclude which is better. The trained models could be used by Information Technology (IT) developers to create mobile apps and websites that can aid in the diagnosis of COVID-19.

The dataset has been processed using balance, augmentation, shuffling, and prefetching techniques described in the methods section. Several neural networks were tested on the datasets using the transfer learning technique. The Google Colab platform was used to run the Python3 code in the form of Jupyter notebooks. The results are explained and discussed in the next sections. The dataset mentioned in Sect. 2.1 was used in this experiment, and we divided it into 3 datasets to test the performance using different class predictions. In Table 4, we have the structure of the dataset used for the 4 classes.

Table 4. Four-class classification arrangement of dataset.

Class	COVID-19	Normal	Viral pneumonia	Lung-opacity	Total
Train	2670	2670	2670	2670	10,680
Validation	572	572	572	572	2288
Test	573	573	573	573	2292
Total	3815	3815	3815	3815	15,260

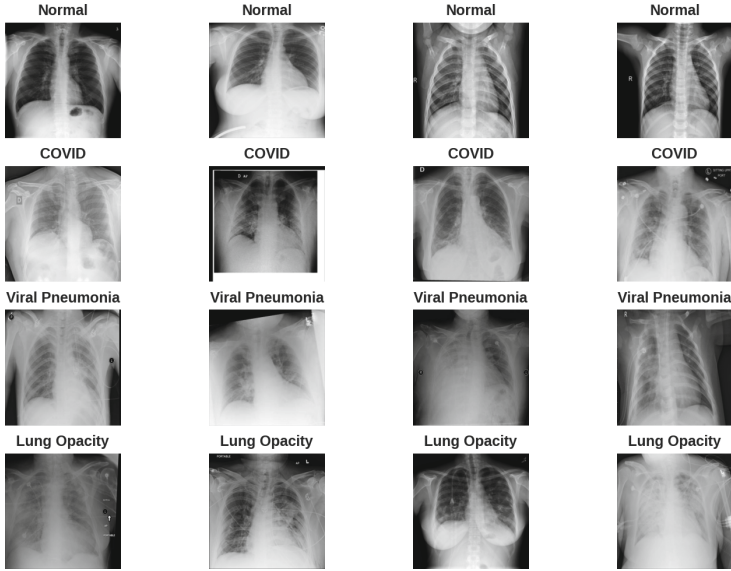


Fig. 4. Samples from the balanced dataset showing all the classes used in the experiment.

In Fig. 4, we can appreciate images from all the 4 classes that were used in this work. The images were resized to reduce GPU ram and training compiling time.

3.1 Transfer Learning Models Architectures

All the transfer learning models have the same general structure: input, sequential, rescaling, functional, pooling, dropout, and dense layers. The functional layer is where the transfer learning model is located. In all the experiments, the dense layer was set to “Trainable,” and the functional layer was frozen during the first process of training. The number of parameters is the same for the freezing and the fine-tuning training. Only the functional and dense layers account for parameters. The details of these layers are reported in Table 5.

Table 5. Transfer learning training parameters. The non-trainable parameters correspond to the functional layer during the freezing process. The functional layer contains the model obtained from the Keras module.

Model architecture	Trainable	Non-trainable	Total
Xception	2049	20,861,480	20,863,529
VGG16	513	14,714,688	14,715,201
VGG19	513	20,024,384	20,024,897

In Fig. 5, we observe the plots produced by the VGG19 model using 4 classes. The freezing training took 32 min and 34 s to compile using T4 GPU over 20 epochs. The fine-tuning stage took 46 min and 16 s to compile over 10 epochs.

3.2 VGG19 Binary Classification Performance

The values to plot the confusion matrix were calculated using the test dataset Fig. 6.

A summary Table 6 is presented to select the best model to be tested on 3 class and 4 class stages. The VGG19 model has the highest accuracy, sensitivity, specificity, precision, and F1 Score. For the sensitivity, it means that the model is better than the others in predicting patients with COVID-19 and has a minimal amount of mispredicted false negatives. Also, for specificity, it means that VGG19 is excellent in predicting patients that do not have the disease and there is a small amount of false positives. In general, for COVID-19 a test that does not miss the true positives would be chosen to avoid the spread of the disease. The F1 Score takes into account both sensitivity and specificity. VGG19 has an F1 Score of 98.53% so it is the best model to be tested in the 3 classes and 4 classes classification.

Xception (a), VGG16 (b), and VGG19 (c) models are shown in Fig. 7. We observe that the AUC are very similar, indicating excellent discrimination, meaning the model excellently distinguishes between the positive and negative COVID-19 classes. We used 45 threshold values for all three graphs: min value: -15.115046 and max value: 13.334197 with an average step: 0.6322. The best-balanced threshold indicates the point where the model will effectively classify the cases. We used Youden's J statistic to determine the best-balanced threshold. The VGG16 has the best graph which might indicate a slightly better performance than the other models. However, using only AUC to determine the best model is not recommended. In our particular experiment, detecting patients positive for COVID-19 is the priority. So, the model that has the highest sensitivity is VGG19. Then we chose VGG19 as the best model to be trained with more classes.

In Fig. 8, the confusion matrix is shown using the 4 classes of the dataset. We observe a lower F1 Score compare to the binary classification.

3.3 Grad-CAM on Our VGG19 Model

In Fig. 9, we observe the Grad-Cam that was obtained using the layer last convolutional layer of our trained VGG19 model named *block5_conv4*. We can appreciate the zones that our model is capturing to predict the class of the image. The red-colored regions are considered highly relevant. Regions highlighted in yellow may still be relevant but to a lesser extent compared to the red regions. Based on the color interpretations, we observe that the left lung has a lot of regions that are considered important. Also, the heart shape is considered as seen in the yellow highlights. These results confirm the annotations mentioned in the state of the art about the reticular interstitial patterns and the appearance of white areas in the lungs called opacities which are areas filled with fluid.

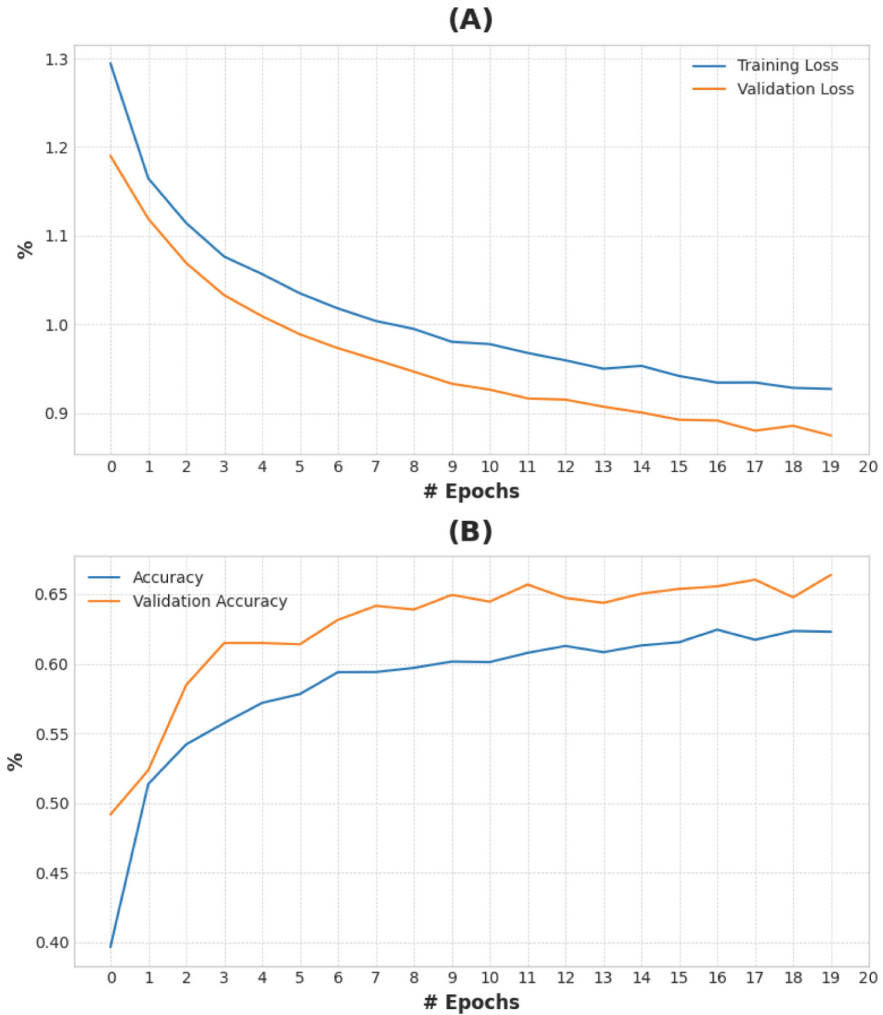


Fig. 5. Convergence plots for the VGG19 model using the classes: Normal, Covid, Viral Pneumonia, and Lung Opacity. **(A) Training loss during freezing:** Both curves converge smoothly because they never overlap each other. **(B) Training freezing accuracy :** The curves present some peaks and again they never overlap, this might indicate that we are going to have more miss-classified images compared to the binary and the 3-class classification. **(C) Loss during fine-tuning:** In this stage we have more loss compared to the 3-class and the binary, the curves converge with almost no peaks. **(D) Fine-tuning accuracy:** Finally, the accuracy is above 90% during this stage, which indicates a good performance.

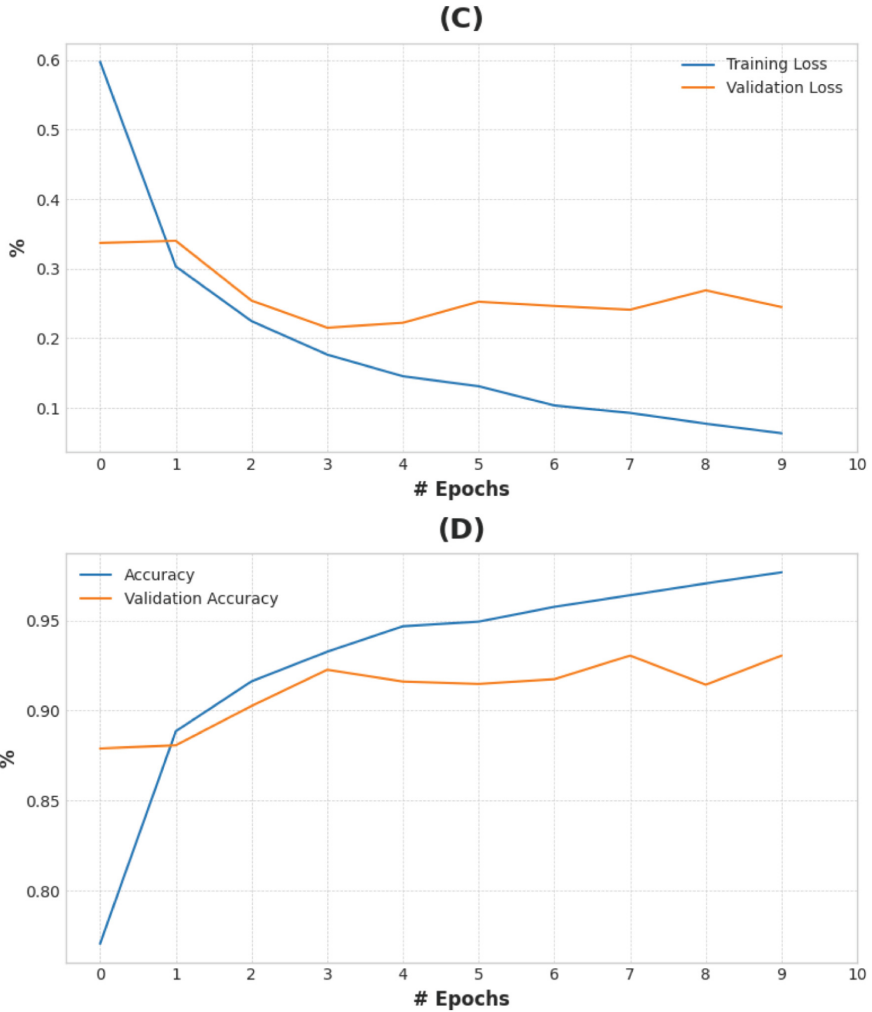


Fig. 5. (*continued*)

In Fig. 10, we observe peripheral bilateral opacities highlighted in red which are very common in COVID-19 patients. So, our trained model is performing excellent in recognizing the patterns of this disease.

4 Discussion of Results

In Table 7, we have summarized the values for accuracy and weighted average F1 score of the VGG19 on the 3 and 4 classes. We observe that adding more classes reduces the model F1 score. The reason might be the similarity in patterns between classes. Overall, the model performed very well on the four-class dataset.

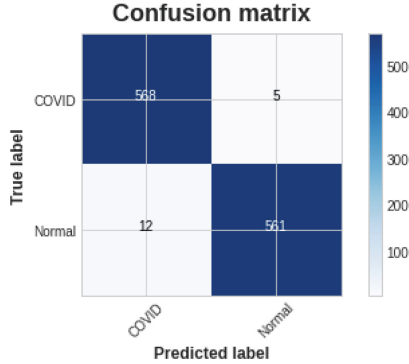


Fig. 6. VGG19 has excellent predictive capability and is slightly better than Xception and VGG16. True Positives: 568, True Negatives: 561, False Positives: 5. False Negatives: 12. Accuracy: 98.52%. Miss-Classification: 1.48%. Sensitivity: 97.93%. Specificity: 99.12%. Precision: 99.13%. F1 Score: 98.53%.

Table 6. Metrics summary for the binary classification using the transfer learning models and COVID and Normal classes.

Model	Accuracy [%]	Sensitivity [%]	Specificity [%]	Precision [%]	F1 score [%]
Xception	95.46	93.34	97.8	97.91	95.57
VGG16	95.72	96.29	95.17	95.11	95.7
VGG19	98.52	97.93	99.12	99.13	98.53

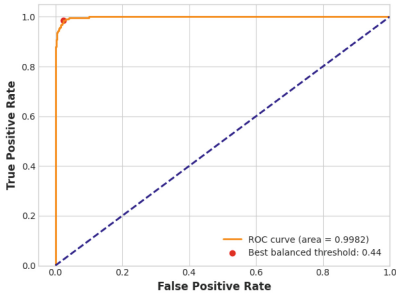
The bold numbers mean the best results

The weighted average F1 score of 89.04% indicates that this model might be used in real applications with very low errors in predicting the image class. The model might be trained on the biggest Kaggle dataset to see if the F1 score improves. Finally, the VGG19 model on the 3 and 4 classes is excellent at predicting the COVID class because the false positive numbers are very low compared to the other classes. We have obtained better results than the work of Sitaula et al. [21] on their VGG19 testing. The reasons might be the use of more images in our work and more pre-processing techniques on the datasets.

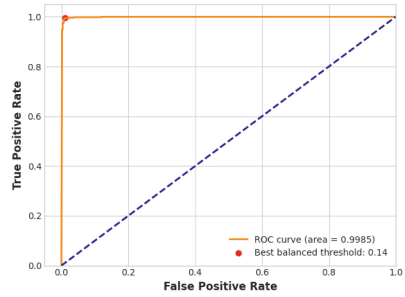
According to the study by [15] they tested modified transfer learning models. They used a data set consisting of 125 COVID-19, 500 atypical pneumonia, and 500 healthy X-ray images. Their data set is very small compared with ours. Let us review the results they obtained in Table 8.

If we compare the VGG19 results of both Tables 6 and 8, our VGG19 transfer learning model performs slightly better with a bigger dataset and 30 epochs compared to the 100 epochs they used. The other modified models they used had very similar results, with Cascade VGGCOV19-NET obtaining the best result.

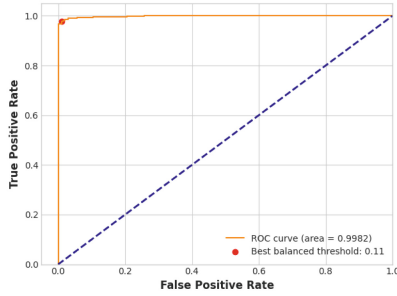
In another study [1], they used 3616 COVID-19 chest X-ray images and 10,192 healthy chest X-rays with VGG19 with 95% accuracy, which is a similar



(a) Xception ROC.



(b) VGG16 ROC.



(c) VGG19 ROC.

Fig. 7. ROC graphs of the models used during binary classification. Roc graphs allow to determine the performance by visual inspection. The more area produced under the curves, the better the model.

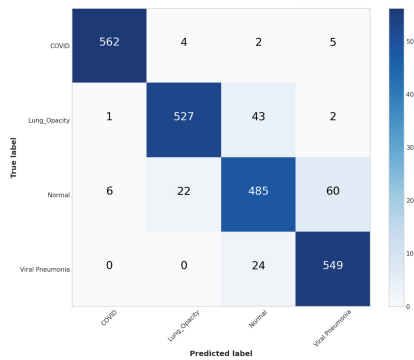


Fig. 8. VGG19 confusion matrix on Normal, Covid, Viral pneumonia, and Lung Opacity classes. Accuracy: 92.63%. Miss-Classification: 7.37%. Weighted average F1 Score: 89.04%.

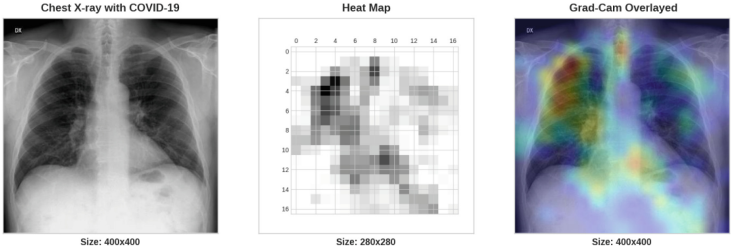


Fig. 9. Chest X-ray of a patient with COVID-19. The first image is extracted randomly from the dataset used in this work. The X-ray image was predicted positive for COVID-19 using the trained model. Next, the Heat Map was obtained using the gradients calculated from the predictions of the model. Then, the Grad-Cam was obtained and overlaid.

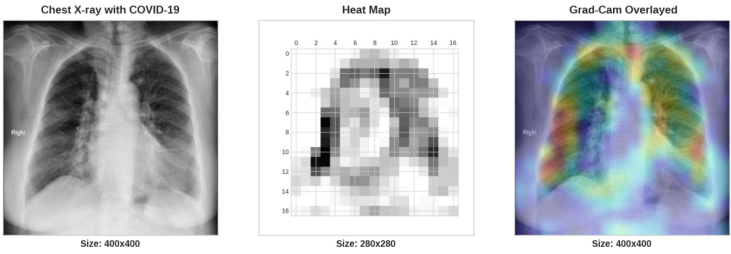


Fig. 10. This image is another example of the Grad-Cam on a patient with COVID-19.

result to our binary accuracy. Now let us review VGG19 in three-class classification.

The worst accuracy in Table 9 is seen in the VGG-19 simple model; the reason might be the lack of use of transfer learning compared to the other studies. If we compare these results with ours, we observe similar results around 92% for the three-class classification.

Table 7. Summary of the metrics that were calculated by using the VGG19 model on three and four-class datasets.

VGG19	Accuracy [%]	Weighted F1 score [%]
3-Classes	92.44	92.32
4-Classes	92.63	89.04

In this work [22], the *COVID-QU-Ex Dataset* was used to test several neural networks using U-net segmentation. The best results are shown in Table 10.

Also, comparing our work results with Table 10 mentioned before, we need to make slight improvements in our experiment. They are using 33,920 images

Table 8. Summary of the results on binary classification of the modified VGG19 CNN.

Model name	F1 score [%]	Accuracy [%]
VGGCOV19-NET	98.00	97.60
Cascade VGGCOV19-NET	99.75	99.84
VGG19	97.68	98.56
Cascade VGG19	98.49	99.04

The bold numbers mean the best results

Table 9. Review of VGG19 transfer learning (TL) models used three class datasets with different sizes.

SourceDataset structure	Method	Accuracy [%]
[3] 224 COVID-19, 700 Pneumonia, 504 Normal	TL-VGG-19	93.48
[8] 1300 COVID-19, 1300 Pneumonia, 1300 Normal	TL-VGG-19	92.92
[11] 130 COVID-19, 140 Pneumonia, 400 Normal	TL-VGG-19	87.00
[20] 260 COVID-19, 300 Pneumonia, 300 Normal	TL-VGG-19	89.30
[6] 1493 COVID-19, 2780 Pneumonia, 1538 Normal	VGG-19	72.52

The bold numbers mean the best results

Table 10. COVID-QU-Ex Dataset tested on state-of-the-art neural networks using segmentation. The U-Net-DenseNet121 and FPN-ResNet18 have the highest sensitivity meaning that they can classify patients with the disease excellently. U-Net++-ResNet18 has the perfect specificity, which means that this neural network classifies, without error, healthy patients. Future work might include replicating this work and trying to achieve similar results using transfer learning with segmentation. They only used three classes, since pneumonia and lung opacity have similar patterns and they merged both classes to get higher performance statistics.

Model	Encoder	F1 score [%]	Accuracy [%]	Sensitivity [%]	Specificity [%]
U-Net	DenseNet121	98.81 ± 0.62	98.8 ± 0.62	99.66 ± 0.33	97.94 ± 0.82
U-Net ++	ResNet18	99.22 ± 0.5	99.23 ± 0.5	98.46 ± 0.71	100 ± 0
U-Net ++	DenseNet121	99.22 ± 0.5	99.23 ± 0.5	99.31 ± 0.48	99.14 ± 0.53
FPN	ResNet18	98.56 ± 0.68	98.54 ± 0.69	99.66 ± 0.33	97.43 ± 0.91

The bold numbers mean the best results

in total in comparison with our 27,052 images. Also, they use segmentation with U-net. Future work might include implementing their large dataset with segmentation and including more neural networks.

4.1 Work Limitations in Our Experiment

The principal limitation we encountered was the lack of computational resources. Initially, the project was tested on a local personal computer with 8GB of RAM and no GPU. The need to use a platform like Google Colab appeared. The compilation of the codes was running at the limit of the crash of the 1 GPU and

16 GB of RAM. When running the multiclass codes, Google Colab crashed due to a lack of RAM space. Then we have to run the codes in Kaggle, which offers a free 30 GB of RAM and two Tesla T4 GPUs. In the future, problems might appear with the compilation of the 33,920 image dataset. So, a paid version might be needed for this future work.

5 Conclusion

In this work, we present Xception, VGG16, and VGG19 convolutional neural networks using the transfer learning technique. The CNNs were trained using the public-access Kaggle dataset that contained chest X-ray images collected before and during the pandemic. The dataset was prepared, separating no relevant archives and not desired images; augmentation was used to increase the number of images and increase the accuracy both for training and prediction; and shuffling allowed for randomness in the training, ensuring better training and variability. We designed the CNNs using the pre-trained models provided by Keras following the documentation for the implementation using transfer learning. We trained, validated, and tested the models using a 70-15-15 percentage ratio. The CNNs successfully classified the images using two, three, and four classes. We used COVID-19, Normal, Lung-Opacity, and Pneumonia classes. The results were reported and compared to other investigations.

The best model in our work was the VGG19. It performed better than the VGG16 and Xception in all the datasets. Our VGG19 neural network obtained a weighted F1 score of 98.53%, 92.32%, and 89.04% for the binary, three, and four class classifications, respectively. We were able to learn what zones of the chest X-ray images is capturing our model by using Grad-Cam. In the discussion of the results we mentioned the work of Tahir et al. that obtained accuracies above 99% using more images with U-net and neural networks. Finally, we stored in Google Drive the trained models which can be deployed to create software tools for public and private hospitals to help diagnose COVID-19.

References




1. Akter, S., Shamrat, F.J.M., Chakraborty, S., Karim, A., Azam, S.: Covid-19 detection using deep learning algorithm on chest X-ray images. *Biology* **10**(11), 1174 (2021)
2. Alshmrani, G.M.M., Ni, Q., Jiang, R., Pervaiz, H., Elshennawy, N.M.: A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alex. Eng. J.* **64**, 923–935 (2023)
3. Apostolopoulos, I.D., Mpesiana, T.A.: Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **43**, 635–640 (2020)
4. Chiu, H.Y., Chao, H.S., Chen, Y.M.: Application of artificial intelligence in lung cancer. *Cancers* **14**(6), 1370 (2022)
5. COVID, K.: Radiography database. Radiological Society of North America (2019). <https://www.kaggle.com/tawsifurrahman/covid19-radiography-database>. Accessed 1 Oct 2021) (19)

6. El Asnaoui, K., Chawki, Y.: Using X-ray images and deep learning for automated detection of coronavirus disease. *J. Biomol. Struct. Dyn.* **39**(10), 3615–3626 (2021)
7. Fatima, N., Liu, L., Hong, S., Ahmed, H.: Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access* **8**, 150360–150376 (2020)
8. Fayemiwo, M.A., Olowookere, T.A., Arekete, S.A., Ogunde, A.O., Odim, M.O., Oguntunde, B.O., Olaniyan, O.O., Ojewumi, T.O., Oyetade, I.S., Aremu, A.A., et al.: Modeling a deep transfer learning framework for the classification of covid-19 radiology dataset. *PeerJ Comput. Sci.* **7**, e614 (2021)
9. Ghoshal, B., Tucker, A.: Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769* (2020)
10. Hansun, S., Argha, A., Liaw, S.T., Celler, B.G., Marks, G.B.: Machine and deep learning for tuberculosis detection on chest X-rays: systematic literature review. *J. Med. Internet Res.* **25**, e43154 (2023)
11. Horry, M.J., Chakraborty, S., Paul, M., Ulhaq, A., Pradhan, B., Saha, M., Shukla, N.: Covid-19 detection through transfer learning using multimodal imaging data. *IEEE Access* **8**, 149808–149824 (2020)
12. INEC: Estadísticas defuncionales (2019). <https://www.ecuadorencifras.gob.ec/documentos/web-inec/Sitios/Defunciones/#ancla-1>. Accessed 30 Sept 2024
13. Jang, M., Kim, M., Bae, S.J., Lee, S.H., Koh, J.M., Kim, N.: Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. *J. Bone Miner. Res.* **37**(2), 369–377 (2022)
14. Juarez, B.: Vgg19 multiclass transfer learning (2023). <https://www.kaggle.com/code/bryanjuarez/vgg19-multiclass-transfer-learning>
15. Karacı, A.: VGGCOV19-net: automatic detection of covid-19 cases from X-ray images using modified VGG19 CNN architecture and yolo algorithm. *Neural Comput. Appl.* **34**(10), 8253–8274 (2022)
16. Mathew, A., Amudha, P., Sivakumari, S.: Deep learning techniques: an overview. *Adv Mach Learn Technol Appl Proc AMLTA* **2020**, 599–608 (2021)
17. Mondal, B.: Artificial intelligence: state of the art. *Recent Trends Adv. Artif. Intell. Internet Things* 389–425 (2020)
18. Organization, W.H.: WHO COVID-19 (2023). <https://covid19.who.int/region/amro/country/ec>. Accessed 30 Sept 2024
19. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of covid-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **121**, 103792 (2020)
20. Rahaman, M.M., et al.: Identification of covid-19 samples from chest X-ray images using deep learning: a comparison of transfer learning approaches. *J. Xray Sci. Technol.* **28**(5), 821–839 (2020)
21. Sitaula, C., Hossain, M.B.: Attention-based VGG-16 model for covid-19 chest X-ray image classification. *Appl. Intell.* **51**, 2850–2863 (2021)
22. Tahir, A.M., et al.: Covid-19 infection localization and severity grading from chest X-ray images. *Comput. Biol. Med.* **139**, 105002 (2021)
23. Tian, M., Shen, Z., Wu, X., Wei, K., Liu, Y.: The application of artificial intelligence in medical diagnostics: a new frontier. *Acad. J. Sci. Technol.* **8**(2), 57–61 (2023)
24. Wang, W., Siau, K.: Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: a review and research agenda. *J. Database Manage.* **30**(1), 61–79 (2019)

25. Widaatalla, Y., Wolswijk, T., Adan, F., Hillen, L., Woodruff, H., Halilaj, I., Ibrahim, A., Lambin, P., Mosterd, K.: The application of artificial intelligence in the detection of basal cell carcinoma: a systematic review. *J. Eur. Acad. Dermatol. Venereol.* **37**(6), 1160–1167 (2023)



Region of Interest Features and Classification of MRI Brain Lesions

Darwin Castillo^{1,3,4,5} , Ricardo J. Alejandro² , Santiago García¹,
María José Rodríguez-Álvarez³ , and Vasudevan Lakshminarayanan^{3,4}

¹ Facultad de Ciencias Exactas y Naturales, Departamento de Química, Universidad Técnica Particular de Loja, 11-01-608, Loja, Ecuador

dpcastillo@utpl.edu.ec

² Department of Experimental Psychology, Ghent University, Ghent, Belgium

³ Instituto de Instrumentación Para Imagen Molecular (i3M), Universitat Politècnica de València – Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain

⁴ Theoretical and Experimental Epistemology Lab, School of Optometry and Vision Science, University of Waterloo, Waterloo, ON, Canada

⁵ Departments of Physics, Electrical and Computer Engineering and Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

Abstract. Nowadays, the diagnosis of numerous diseases is facilitated by medical imaging. In that context, the identification of brain lesions presented as White Matter Hyperintensities (WHMs) and their related diseases is essential to have a correct diagnosis. Machine- and deep learning (subfields within artificial intelligence) could support the diagnosis (especially in complex medical images) by leveraging the structure and regularities within the imaging data. This project presents a technique for the classification of WHMs concerning ischemia and demyelination through the analysis of the region of interest (ROI) features of magnetic resonance images. To do that, we analyzed radiomic features using a combination of principal component analysis (PCA) and support vector machine (SVM) classification. Next, we used a transfer learning fine-tuned ResNet18 model to more thoroughly analyze and classify lesioned ROIs. For that, we used patient data alone and additional synthetic data (generated using spectral generative adversarial networks -SNGAN). The results show an accuracy mean value of 0.96 without data augmentation; while we had a value of 0.54 using synthetic data, a similar value was acquired with radiomics-informed SVM classification (0.56). These findings constitute a starting point for future projects exploring different ways of informing and fine-tuning artificial intelligence models to detect, classify, and segment MRI pathologies characterized by small lesions.

Keywords: ROIs · brain lesions · radiomics · GAN · resnet18 · machine learning · medical images

1 Introduction

Medical imaging is the primary tool for the rapid and non-invasive diagnosis of many types of cardiovascular and respiratory diseases, cancer, and diabetes [1]. The diagnosis of neurological pathologies, in particular, has been largely facilitated, given the particular

constraints for access and sensitivity to damage that the human brain presents [2], in the sense that lesions/infarctions in other body systems might lead to minor symptoms, but when present in the brain, they often lead to significant impairment.

Two significant instances of brain dysfunction are stroke and demyelinating diseases (e.g., multiple sclerosis). Stroke (or ischemic stroke, more specifically) is one of the leading causes of mortality and long-term disability in middle- and high-income countries [3], characterized by an inadequate blood supply that leads to cell death [2]. Multiple Sclerosis (MS), a chronic neurodegenerative disease, is mainly characterized by the loss of myelin (the protective covering of nerves, including the brain and spinal cord) that leads to progressive and diverse neurological deficits [4].

The brain lesions produced by both ischemic stroke and demyelinating appear as white matter hyperintensities (WMHs) [5, 6] in T2-weighted, diffusion-weighted (DW), and fluid-attenuated inversion recovery-(FLAIR) magnetic resonance imaging (MRI) [7]. The MRI similarities between the pathologies make it difficult to differentiate between them (and could potentially lead to misdiagnoses) without proper training [8, 9]. Hence, these lesions are generally detected by experienced radiologists.

Machine learning (ML) and deep learning (DL) are two subfields within artificial intelligence that leverage the latent structures and relations across data samples to characterize the nature of entire datasets. In that sense, artificial intelligence could aid the diagnosis of ischemic and demyelinating diseases by learning the MRI-based specific attributes to identify each pathology correctly (and automatically).

In this sense, this project outlines the analysis of the WMH brain lesions due to ischemic stroke and demyelinating lesions from the radiomics view, exploring the features of the lesions and, afterward, a deep learning methodology for the classification using a transfer learning process.

From a radiomics view, we conducted a detailed study of the lesions by extracting radiomic features [10], followed by multidimensional scaling (MDS) through Principal Component Analysis (PCA) [11] to identify pathology-related features to inform a machine learning classifier.

From a deep learning view, we implemented a classification algorithm leveraging a transfer learning process. Specifically, we used a fine-tuned ResNet18 network [12], pre-trained on the ImageNet database [13]. Additionally, to extend our dataset and, by so doing, to improve the predictive capabilities of our model, we used spectral normalization generative adversarial networks (SNGAN) for synthetic data generation (data augmentation) [14].

1.1 Related Work

Several studies [5, 6, 15–17] have proposed different methods and ML/DL techniques for the segmentation, detection, and classification of brain lesions concerning WMHs. For instance, Mitra et al. [16], using a Random Forest (RF) classifier, reported a Dice score (DSC) of 0.60 between stroke and MS. Ghaforian et al. [17] reported a sensitivity of 0.73 using a combination of AdaBoost and RF algorithms to detect cerebral small-vessel disease.

The use of clustering and variance detection algorithms has shown promise for neuroimaging disease classification. Principal component analysis, in particular, has

been used (among other cases) to improve the detection capabilities of neural networks [18] and as a feature selector to improve/direct classification accuracy in perceptron-based models [19]. Such evidence tells us that leveraging high-dimensional data by identifying their main features is a promising avenue to explore.

On the other hand, the advent of neural networks has provided enhanced detection and classification algorithms. Guerrero et al. [15] presented DSC values of 69.5 through the use of CNNs with a u-shaped residual network architecture (uResNet) to differentiate the disease-specific WMHs. Along the same line, [20] mentions that the ResNet18 network, due to its optimization and efficiency with respect to training time, feature extraction, and transfer learning implementation, shows broad applicability for the classification and identification of pathologies in medical images, e.g., intracranial diagnosis [21], skin cancer [22], breast cancer [23, 24], brain tumors [25–27], stroke [3, 28], COVID-19 detection [29].

A study [30] using ResNet18 and ResNet50 reported accuracy values of 96.6 and 98.45 to classify breast cancer when focusing on the details of ROIs. In [31], the expansion of ROIs showed an improved classification performance of brain tumors. For that matter, fine-tuned ResNet18, GoogLeNet, and other neural networks were used as feature extractors.

The amount (and diversity) of data has a central role in the training of ML and DL algorithms in order to improve the accuracy of the models. However, medical images tend to lack numerous observations per condition in datasets. In that sense, generative adversarial networks (GAN) [14] constitute a tool to increase data samples and expand existing medical image datasets due to their capabilities to generate synthetic images [32]. An example of this application is presented in [33], where a framework called BrainGAN was proposed to generate and classify brain MRI images using GAN architectures and DL models.

2 Materials and Methods

2.1 Data

A private Hospital in Ecuador proportionated the dataset used in this project, composed of 220 images of the FLAIR modality of MRI [34]. The images were collected using a 1.5 T Philips Achieva medical MRI system.

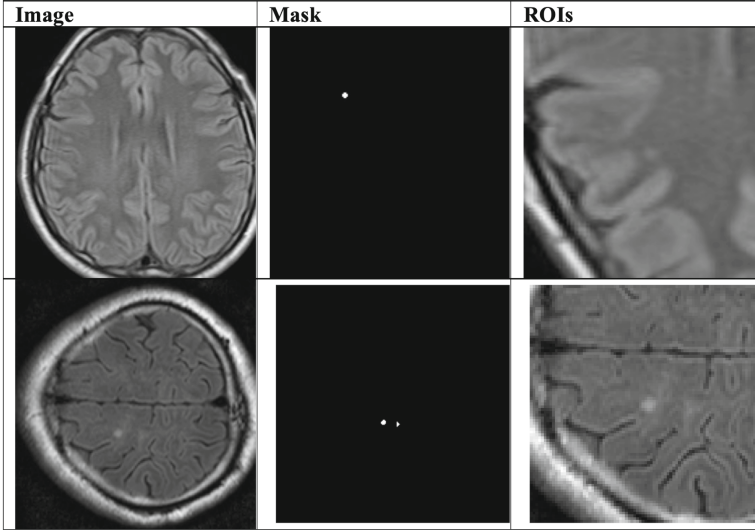
Deep learning studies have shown that for medical imaging analysis, it is essential to extract the region of interest (ROI; the specific area within an image that shows the feature of interest) to train an accurate lesion classifier [30]. Hence, we first extracted lesion-related ROIs from the preprocessed images.

The process for selecting ROIs is shown in Fig. 1. The number of ROIs extracted from the 220 FLAIR images is shown in Table 1, which also presents the proportion of the dataset destined for training (80%) and validation (20%). Additionally, synthetic data was generated with these ROIs to increase the dataset by 4000 additional ROIs.

The ROIs for the classifier were extracted by correlating the images with their lesions' binary masks. After overlapping the images with their mask, the region is cropped accordingly, resulting in a new image (size = 128×128 pixels).

Table 1. Number of extracted and generated ROIs

	Demyelination	Ischemia	Total
	260	261	521
Training (80%)	201	215	416
Validation (20%)	59	46	115
Synthetic ROIs generated	2000	2000	4000

**Fig. 1.** Process of generation of ROIs for the classifier

2.2 Image Features and Principal Component Analysis

To characterize and potentially uncover any latent features present in the MRI data that might facilitate direct pathology classification (ischemia versus demyelination), we used the open-source package PyRadiomics [35, 36] to extract the data’s radiomic features.

As output, PyRadiomics provides around 1500 features (categorized under first-order statistics, shape descriptors, gray level properties, etc.; for a full description, see [35]) per image that have proven to be efficient predictors in automated diagnosis by medical imaging [35, 36] (Fig. 2 and 3).

Given the vast search space within each feature set per image, we used the scikit-learn module [37] to perform PCA to reduce the feature dimensionality space and retained the main two and three components (see Fig. 4). The rationale of this approach being that if these pathologies can be differentiated by their radiomic features, then the principal components (PCs) explaining most of the variance (>80%) should contain the necessary information that allows for such discrimination. The PCA algorithm used

single-value decomposition to project our complex high-dimensional data into lower-dimensional spaces [38] while preserving the pairwise distances between data points (measured as Euclidean distances) as much as possible.

Subsequently, we performed a cross-validated classification analysis. On each fold of the cross-validation procedure, the data was split into 75 and 25% for training and testing, respectively, to go into a support vector machine (SVM) classifier (with a radial basis function kernel and regularization = 1.0). Then, the aggregate classification accuracy across folds would be deemed as the model classification accuracy.

2.3 GAN Network Architecture

The generation of synthetic data used the SNGAN network. The model utilized standard convolutional layers and LeakyReLU (with batch normalization) as the activation function for the generator and transposed layers with spectral normalization for the discriminator. This architecture enables the generation of images of sizes 128×128 px.

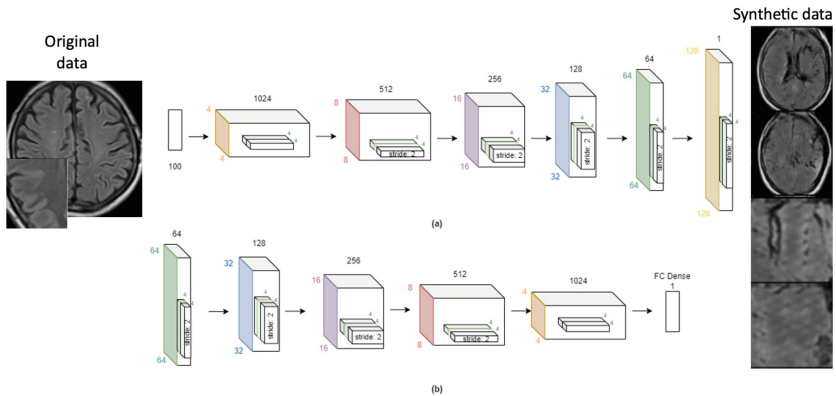


Fig. 2. The architecture of the Generator (a) and Discriminator (b) of the SNGAN model proposed for ROIs data synthetic image generation. Figure adapted from [39]

Table 2 shows the hyperparameter used to generate the synthetic images through different experiments.

Table 2. Hyperparameter for data generation

Size image	128×128	Epochs	400
Batch size	32	Optimizer	Adam
Loss	Binary cross entropy with logistic loss	Learning rate (lr)	0.00015

2.4 Model Training

The ResNet18 model is formed by five blocks (see Fig. 3), each containing a convolution identity block, where both the convolution and identity blocks consist of three convolutional layers connected with skip connections, enabling forward and backward propagation. The model receives input images resized to 128×128 pixels, and the network is designed for binary classification.

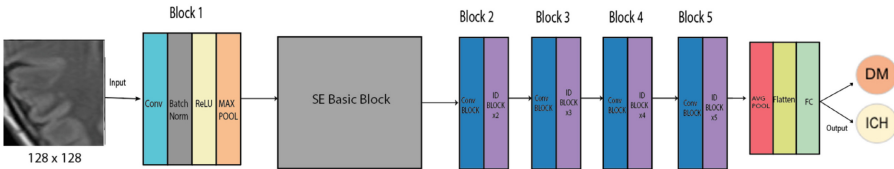


Fig. 3. The ResNet18 model was used to classify the ROI lesions of ischemia (ICH) and demyelination (DM).

This approach effectively mitigates the issues of vanishing or exploding gradients that can arise from increasing neural network depth, ultimately leading to improved accuracy [40]. Table 3 shows the ResNet18 model hyperparameters used to train the classification.

Table 3. Hyperparameter for data classification model using ResNet18

Image size	128×128	Epochs	100
Batch size	16	Optimizer	Adam
Loss	Binary cross entropy with logistic loss	Learning rate (lr)	0.0001

All the experiments in this project were done using the Pytorch framework and were run using the Google Colab version Pro Tesla V100-SXM2-16GB. The time per epoch in the generation of synthetic data had an average value of 64 s, and for the Classification model using the Resne18 model, it was 108 s per epoch. Data generation required more raining epochs (400) than classification 100.

3 Results

3.1 Principal Component Analysis and Feature Classification

Figure 4 depicts principal two (panel a) and three (panel b) components retained after dimensionality reduction (of a radiomics space of ~ 1500 components). The classification of these components into pathologies (ischemia vs. demyelination) obtained an average accuracy of 0.56, suggesting that although the main components provide qualitatively distinct pathology radiomic features, their quantitative differences do not allow for completely efficient discrimination between the two brain lesions.

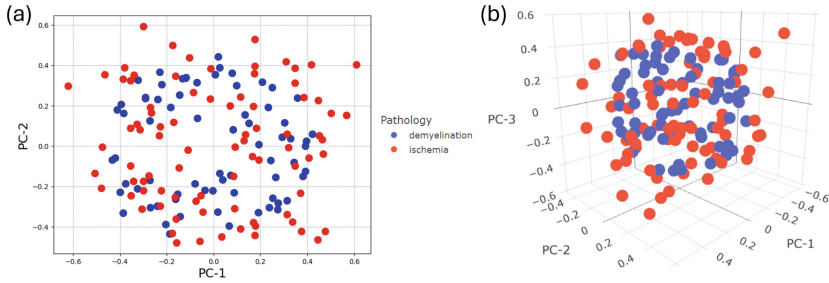


Fig. 4. Visualization of the main (a) 2 and (b) 3 principal components (PC) resulting from principal component analysis (PCA).

3.2 ResNe18 Model Classification

Table 4 describes the metrics classification of the ROIs model with the data without data augmentation. These results were acquired using the following hyperparameters: batch size = 16, $\beta_1 = 0.1$, $\beta_2 = 0.75$, $lr = 0.0001$. Figure 5 shows the curves of the classification metrics and the curves of loss values.

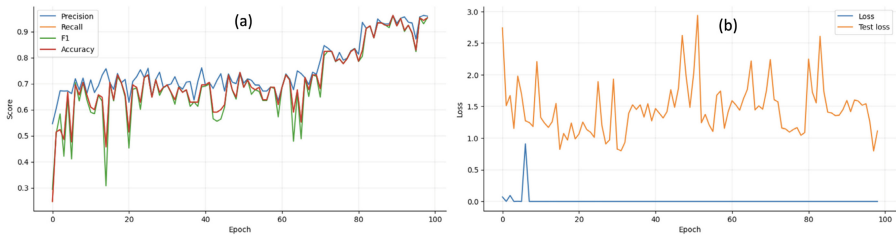


Fig. 5. Graphs of the evolution of (a) classification metrics and (b) Loss values of the model without synthetic data during the training.

Figure 6 shows the Confusion Matrix of the validation data. Table 4 shows the classification metrics of the ResNet18 model without data augmentation. We can see (in Fig. 6 and Table 4) that the model is increasingly accurate in classifying the types of lesions.

Table 4. Classification metrics of the classification model with data without data augmentation

Lesion/Metric	Accuracy	Precision	Recall	F1-score
Demyelination	94.9%	91%	88%	90%
Ischemia	97.8%	87%	91%	89%

		Model Predictions	
		Demyelination	Ischemia
Lesions	Demyelination	56	3
	Ischemia	1	45

Fig. 6. Confusion Matrix of validation data using the ROIs classification model without the data augmentation. Numbers represent the number of images being classified.

3.3 ResNet18 Model Classification with Synthetically Generated Data:

Through the SNGAN model with the following hyperparameters: batch size = 16, $\beta_1 = 0.1$, $\beta_2 = 0.99$, $lr = 0.0015$, we generated synthetic images to increase the sample size of the dataset.

The Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) metrics were used to evaluate the synthetic data generated. These metrics allow us to determine the difference in vector representation between the synthetic and real images.

FID permits the comparison of the distributions of the original and synthetic images. Better-quality images are indicated by lower FID scores [41]. KID shows the degree of visual similarity between the generated and real images. A lower KID value means a high quality of visual similarity. In general, low FID and KID values denote good results for synthetic image generation. Details of the KID and FID equations are shown in [23, 41, 42].

However, there are no benchmark values since they depend on the context of generation, especially in medical images [43, 44] (we refer the reader to [44, 45] for a comparison and repository of different values in experiments generating medical images).

Through an analysis of the curves in Fig. 7, we can conclude:

- The Discriminator identifies real images, stabilizing around 80–100%.
- The accuracy of identifying fake images indicates that the generator has gradually improved in generating more realistic images.
- The Generator Loss shows a decreasing trend with significant fluctuations, indicating the generator’s ongoing efforts to improve.
- The Discriminator Loss stabilizes at a low value, suggesting that the discriminator effectively distinguishes between real and fake images.
- The FID value 340 indicates an improvement in the generated images’ quality as training progresses. However, as we can see in Fig. 7, the images are not as high-quality as the original images.
- The KID values stabilize at a value of 0.38, which suggests the model improves in generating realistic images.

The fluctuations in graphs are typical in GAN training due to the adversarial setup. However, we can see a stabilization in both the discriminator accuracy and the FID/KID scores in the latter epochs, suggesting a convergence toward better performance.

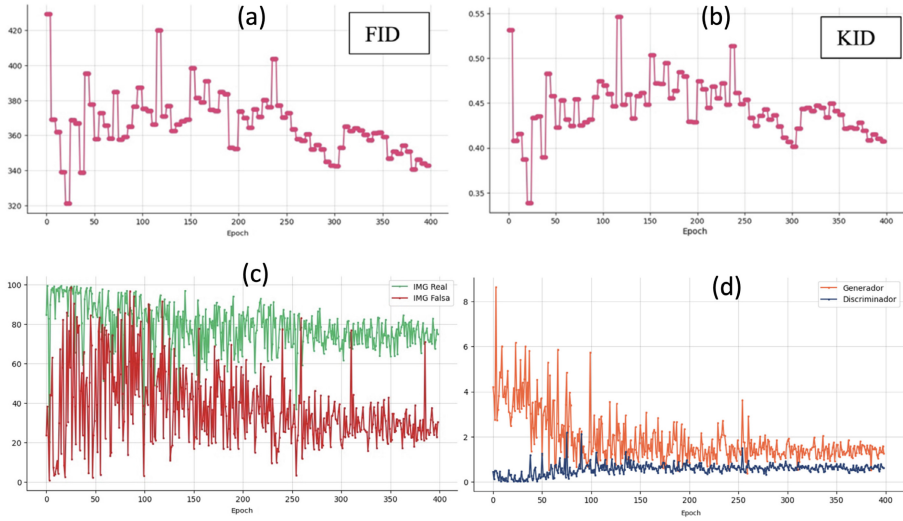


Fig. 7. Curves of (a) Fréchet inception distance (FID) with a mean value of 342.57, and (b) Kernel Inception Distance (KID) with a mean value of 0.401. (c) Discriminator accuracy. (d) Generator and Discriminator Losses of the synthetic data generated over 400 epochs for model SNGAN

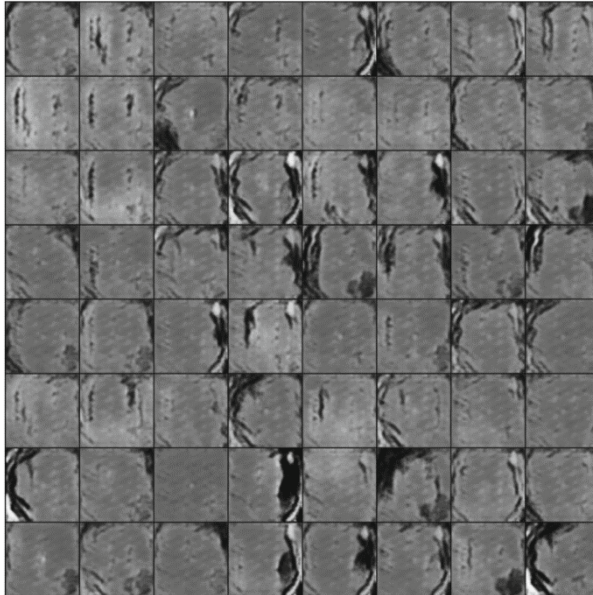


Fig. 8. Example of the synthetic data generated with the best SNGAN model (400 epochs).

Table 5 describes the metrics classification of the ROIs model with synthetic data. These results were acquired using the following hyperparameters: batch size = 16,

$\beta_1 = 0.1, \beta_2 = 0.75, lr = 0.0001$. Figure 9 shows the curves of the classification metrics and the curves of loss values. Figure 10 shows the confusion matrix of validation data of the ROI classification using data augmentation.

In this case, Table 5 shows that the mean value of classification accuracy of the model is 54% for the two categories.

With these values in mind, we can see that the image generation model allows the construction of an extended training dataset for training. However, to have a better classification performance, it is necessary to improve the quality of these synthetic data (see Fig. 8).

Table 5. Classification metrics of the classification model with data augmentation

Lesion/Metric	Accuracy	Precision	Recall	F1-score
Demyelination	55.9%	51%	48%	50%
Ischemia	51.1%	51%	54%	53%

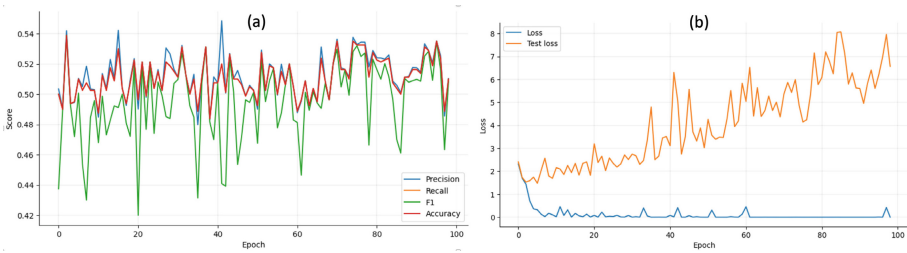


Fig. 9. Graphs of Classification metrics (a) and Loss values (b) of the model using the synthetic data generated

			Model Predictions	
			Demyelination	Ischemia
Lesions	Demyelination	192	207	
	Ischemia	184	217	

Fig. 10. Confusion Matrix for validation data of the ROIs classification model using the synthetic data generated

4 Brief Discussion and Conclusions

Training a classifier with only ROIs allowed us to achieve a better classification of lesions that are not well delineated, as well as small lesions. In this work, the classifier produced a satisfactory classification result (0.96), similar to that reported in the literature by [30,

31] for the classification of breast cancer and brain tumors, respectively. We attribute such improvement in accuracy to the model being able to explicitly focus on learning the features of the lesions.

Before implementing deep learning-based classification models, a detailed study of the lesions was conducted using radiomic features, texture analysis, and PCA multidimensionality scaling. Support vector classification (ischemia versus demyelination) of these features reached an accuracy of 0.56, suggesting that the radiomic properties in our dataset are highly similar across conditions to establish quantitative differences between them from radiomic features alone.

One reason for the high similarity in image features across pathologies could be the lesion size per image. With only a few pixels per image (less than 0.4 pixels and less than one voxel in 3D series), it would not be surprising to reach a chance-level classification accuracy. Although the performance of our SVM classifier exceeded random noise, the reliability of the predictions needs improvement. Nevertheless, our radiomics analysis provides an encouraging starting point for further investigating this kind of classification; for instance, a richer dataset might provide additional statistical power and sample type heterogeneity to promote the intended discrimination.

The deep learning-based classification experiments without data augmentation gave an average value of accuracy of 0.96, while synthetically generated data of ROIs gave a value of 0.54 in the classification.

One of the reasons for these low values with the synthetic data is because, as noted in [46], the quality of generated medical images is different than natural images, and in that sense, there are some limitations, such as the interpretability and the high computational cost of model training [47, 48].

It is important to note that this work primarily focuses on the performance of lesion classification and, therefore, requires before the segmentation of masks. In future work, this step will be enhanced by integrating an algorithm like the Segment Anything Model (SAM), enabling automatic segmentation of the region of interest (ROI). Additionally, this approach will allow physicians to select the ROI manually and subsequently view the classification results. However, we expect the models developed in this study to become a starting point for upcoming research in the frontier of utilizing GANs in medical imaging and the classification (and segmentation) of small lesions.

The code and data generated in this project will be available upon reasonable request at <https://github.com/aiformedicine/mriclass.git>.

Acknowledgment. The authors acknowledge the research support to Universidad Técnica Particular de Loja through the project PROY_INV_QU_2022_3576. D.C. also acknowledges the support from Universitat Politècnica de València through Assistance Call Doctoral Student Mobility. R.J.A. is supported by funding from the European Research Council under the EU's Horizon 2020 Research and Innovation Programme.

References

1. Hussain, S., et al.: Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *Biomed. Res. Int.* **2022**, 5164970 (2022). <https://doi.org/10.1155/2022/5164970>



2. Favate, A.S., Younger, D.S.: Epidemiology of ischemic stroke. *Neurol. Clin.* **34**, 967–980 (2016). <https://doi.org/10.1016/J.NCL.2016.06.013>
3. Fang, G., Huang, Z., Wang, Z.: Predicting ischemic stroke outcome using deep learning approaches. *Front Genet.* **12**, (2022). <https://doi.org/10.3389/FGENE.2021.827522/FULL>
4. Thompson, A.J., Baranzini, S.E., Geurts, J., Hemmer, B., Ciccarelli, O.: Multiple sclerosis. *The Lancet* **391**, 1622–1636 (2018). [https://doi.org/10.1016/S0140-6736\(18\)30481-1](https://doi.org/10.1016/S0140-6736(18)30481-1)
5. Karthik, R., Menaka, R., Hariharan, M., Won, D.: Ischemic lesion segmentation using ensemble of multi-scale region aligned CNN. *Comput. Methods Programs Biomed.* **200**, 105831 (2021). <https://doi.org/10.1016/J.CMPB.2020.105831>
6. Park, G., Hong, J., Duffy, B.A., Lee, J.M., Kim, H.: White matter hyperintensities segmentation using the ensemble U-Net with multi-scale highlighting foregrounds. *Neuroimage* **237** (2021). <https://doi.org/10.1016/J.NEUROIMAGE.2021.118140>
7. Zoghi, A., Petramfar, P., Nikseresht, A., Sakhaee, E.: Investigation of ischemic and demyelinating lesions by cerebral vasoreactivity based on transcranial Doppler sonography: a comparative study. *Neuropsychiatr. Dis. Treat.* **14**, 2323–2328 (2018). <https://doi.org/10.2147/NDT.S150062>
8. Castillo, D., Lakshminarayanan, V., Rodríguez-Álvarez, M.J.: Mr images, brain lesions, and deep learning. *Appl. Sci. (Switzerland)* **11**, 1–41 (2021). <https://doi.org/10.3390/app11041675>
9. Castillo, D.P., Samaniego, R.J., Jimenez, Y., Cuenca, L.A., Vivanco, O.A., Alvarez-Gomez, J.M., Rodriguez-Alvarez, M.J.: Identifying Demyelinating and Ischemia brain diseases through magnetic resonance images processing. In: *IEEE Nuclear Science Symposium, Medical Imaging Conference, NSS/MIC. Institute of Electrical and Electronics Engineers Inc.* (2019). <https://doi.org/10.1109/NSS/MIC42101.2019.9060022>
10. Yip, S.S.F., Aerts, H.J.W.L.: Applications and limitations of radiomics. *Phys. Med. Biol.* **61**, R150 (2016). <https://doi.org/10.1088/0031-9155/61/13/R150>
11. Greenacre, M., Groenen, P.J.F., Hastie, T., D’Enza, A.I., Markos, A., Tuzhilina, E.: Principal component analysis. *Nature Rev. Methods Primers* **2**(1), 2, 1–21 (2022). <https://doi.org/10.1038/s43586-022-00184-w>
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *He_2016_CVPR* (ed.) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016)
13. ImageNet: <https://www.image-net.org/update-mar-11-2021.php>. Last accessed 2024/07/12
14. Goodfellow, I., et al.: Generative adversarial nets. *Adv. Neural. Inf. Process. Syst.* **27**, 2672–2680 (2014)
15. Guerrero, R., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *Neuroimage Clin.* **17**, 918–934 (2018). <https://doi.org/10.1016/j.nicl.2017.12.022>
16. Mitra, J., et al.: Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *Neuroimage* **98**, 324–335 (2014). <https://doi.org/10.1016/J.NEUROIMAGE.2014.04.056>
17. Ghafoorian, M., et al.: Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Med. Phys.* **43**, 6246–6458 (2016). <https://doi.org/10.1118/1.4966029>
18. Ahmad, I., Siddiqi, M.H., Alhujaili, S.F., Alrowaili, Z.A.: Improving Alzheimer’s disease classification in brain MRI images using a neural network model enhanced with PCA and SWLDA. *Healthcare* **11**, 2551 (2023). <https://doi.org/10.3390/HEALTHCARE11182551>
19. Olle Olle, D.G., Zoobo Bisse, J., Abessolo Alo’o, G.: Application and comparison of K-means and PCA based segmentation models for Alzheimer disease detection using MRI. *Discov. Artif. Intell.* **4**, 1–14 (2024). <https://doi.org/10.1007/S44163-024-00106-7/FIGURE S/11>

20. Zhao, Y., Zhang, X., Feng, W., Xu, J.: Deep learning classification by ResNet-18 based on the real spectral dataset from multispectral remote sensing images. *Remote Sens.* **14**, 4883 (2022). <https://doi.org/10.3390/RS14194883>
21. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J.: Deep learning for brain MRI segmentation: state of the art and future directions. **30**, 449–459 (2017)
22. Gouda, N., Amudha, J.: Skin cancer classification using ResNet. In: 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020. 536–541 (2020). <https://doi.org/10.1109/ICCCA49541.2020.9250855>
23. Jiménez-Gaona, Y., et al.: BraNet: a mobil application for breast image classification based on deep learning algorithms. *Med. Biol. Eng. Comput.* (2024). <https://doi.org/10.1007/S11517-024-03084-1>
24. Mu, J.: Breast cancer detection using ResNet with Hyperparameter tuning. In: 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications, ICPECA 2023, pp. 1245–1249 (2023). <https://doi.org/10.1109/ICPECA56706.2023.10076197>
25. Khodadadi Shoushtari, F., Sina, S., Dehkordi, A.N.V.: Automatic segmentation of glioblastoma multiform brain tumor in MRI images: using Deeplabv3+ with pre-trained Resnet18 weights. *Physica Med.* **100**, 51–63 (2022). <https://doi.org/10.1016/J.EJMP.2022.06.007>
26. Jiao, T., et al.: Deep learning with an attention mechanism for differentiating the origin of brain metastasis using MR images. *J. Magn. Reson. Imaging* **58**, 1624–1635 (2023). <https://doi.org/10.1002/JMRI.28695>
27. Kibriya, H., Masood, M., Nawaz, M., Rafique, R., Rehman, S.: Multiclass brain tumor classification using convolutional neural network and support vector machine. In: Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing, MAJICC 2021. (2021). <https://doi.org/10.1109/MAJICC53071.2021.9526262>
28. Sirsat, M.S., Fermé, E., Câmara, J.: Machine learning for brain stroke: a review. *J. Stroke Cerebrovasc. Dis.* **29**, 105162 (2020). <https://doi.org/10.1016/J.JSTROKECEREBROVADIS.2020.105162>
29. Showkat, S., Qureshi, S.: Efficacy of transfer learning-based ResNet models in chest X-ray image classification for detecting COVID-19 Pneumonia. *Chemom. Intell. Lab. Syst.* **224**, 104534 (2022). <https://doi.org/10.1016/J.CHEMOLAB.2022.104534>
30. Sinha, A., Naskar, M.N.B.J., Pandey, M., Rautaray, S.S.: ROI Segmentation for breast cancer classification: deep learning perspective. In: 2023 IEEE 4th Annual Flagship India Council International Subsections Conference: Computational Intelligence and Learning Systems, INDISCON 2023 (2023). <https://doi.org/10.1109/INDISCON58499.2023.10269829>
31. Öksüz, C., Urhan, O., Güllü, M.K.: Brain tumor classification using the fused features extracted from expanded tumor region. *Biomed. Signal Process. Control* **72**, 103356 (2022). <https://doi.org/10.1016/J.BSPC.2021.103356>
32. Tavse, S., Varadarajan, V., Bachute, M., Gite, S., Kotecha, K.: A systematic literature review on applications of GAN-synthesized images for brain MRI. *Future Internet* 2022, **14**, 351 (2022). <https://doi.org/10.3390/FI14120351>
33. Alrashedy, H.H.N., Almansour, A.F., Ibrahim, D.M., Hammoudeh, M.A.A.: BrainGAN: brain MRI image generation and classification framework using GAN architectures and CNN models. *Sensors (Basel)* **22** (2022). <https://doi.org/10.3390/S22114297>
34. Kamr, W.H., Almalki, S., Saadawy, A.M.I., El-Tahan, A.: FLAIR vascular hyperintensity, an early sign of stroke (case report). <https://doi.org/10.1186/s43055-021-00627-6>
35. Zwanenburg, A., et al.: The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020). <https://doi.org/10.1148/RADIOL.2020191145/ASSET/IMAGES/LARGE/RAD IOL.2020191145.FIG5.JPEG>

36. Van Griethuysen, J.J.M., et al.: Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017). <https://doi.org/10.1158/0008-5472.CAN-17-0339/SUPPLEMENTARY-VIDEO-S2>
37. Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., et al.: Scikit-learn: machine learning in Python. *Matthieu Perrot. J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
38. Szlam, A., Kluger, Y., Tygert, M.: An implementation of a randomized algorithm for principal component analysis. *ACM Trans. Math. Softw.* **1**, 1, Article 1, (2014). <https://doi.org/10.1145/0000000.0000000>
39. García, S.: Aumento De Imágenes De Mama Mediante Redes Generativas Antagónicas (gans). <http://dspace.utpl.edu.ec/jspui/handle/123456789/37590> (2023)
40. Joshi, D., Channe, H.: A survey on brain tumor detection based on structural MRI using machine learning and deep learning techniques. *Int. J. Sci. Technol. Res.* **9**, 17–23 (2020)
41. Borji, A.: Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.* **179**, 41–65 (2019). <https://doi.org/10.1016/J.CVIU.2018.10.009>
42. Parmar, G., Zhang, R., Zhu, J.Y.: On Aliased resizing and surprising Subtleties in GAN evaluation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2022-June*, 11400–11410 (2021). <https://doi.org/10.1109/CVPR52688.2022.01112>
43. Abdusalomov, A.B., Nasimov, R., Nasimova, N., Muminov, B., Whangbo, T.K.: Evaluating synthetic medical images using artificial intelligence with the GAN algorithm. *Sensors (Basel)* **23** (2023). <https://doi.org/10.3390/S23073440>
44. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarbuerger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J.N., Truhn, D.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci. Rep.* **13**, (2023). <https://doi.org/10.1038/S41598-023-39278-0>
45. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z.A., Yang, Y.: RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* **4** (2022). <https://doi.org/10.1148/RYAI.210315>
46. Skandarani, Y., Jodoin, P.M., Lalande, A.: GANs for medical image synthesis: an empirical study. *J Imaging* **9**, (2023). <https://doi.org/10.3390/JIMAGING9030069/S1>
47. Chen, Y., et al.: Generative adversarial networks in medical image augmentation: a review. *Comput. Biol. Med.* **144**, 105382 (2022). <https://doi.org/10.1016/J.COMPBIOMED.2022.105382>
48. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: *4th International Conference on Learning Representations, ICLR 2016—Conference Track Proceedings* (2016)



Deep Learning-Based Leukemia Diagnosis from Bone Marrow Images

Luis Zhinin-Vera^{1,2,3} , Alejandro Moya¹ , Elena Pretel¹ ,
Jaime Astudillo² , and Javier Jiménez-Ruescas¹ 

¹ LoUISE Research Group, University of Castilla-La Mancha, 02071 Albacete, Spain

² School of Mathematical and Computational Sciences, Yachay Tech University, 100650 Urcuquí, Ecuador

³ MIND Research Group—Model Intelligent Networks Development, Urcuquí, Ecuador

luis.zhinin@uclm.es

Abstract. Identifying and classifying features in Bone Marrow Aspirate Smear (BMAS) images is essential for diagnosing various leukemias, such as Acute Myeloid Leukemia. The complexity of microscopy image analysis necessitates a computational tool to automate this process, reducing the workload on hematologists. Our study introduces a Deep Learning-based method designed to efficiently detect and classify cell characteristics in BMAS images. Current systems struggle with cell and nucleus segmentation due to variations in cell size, appearance, texture, and overlapping cells, often influenced by different microscopy conditions. We addressed these challenges by experimenting with the Munich AML Morphology Dataset and a custom dataset from Hospital 12 de Octubre in Madrid. The proposed system achieved over 90% accuracy and 92% precision in identifying and classifying leukemia cells, marking a substantial advancement in supporting clinical specialists in their decision-making processes when traditional analysis methods are insufficient.

Keywords: deep learning · image classification · leukemia cells · bone marrow aspirate smear

1 Introduction

Leukemia is a severe blood disorder characterized by the uncontrolled proliferation of abnormal white blood cells, posing significant health challenges [15]. One of the approaches to tackle this problem is through the use of medical imaging. Advances in digital scanning and camera-integrated microscopes have revolutionized the way cell and tissue images are stored and preserved, which is pivotal for ongoing research into various diseases [22]. These technologies safeguard samples from deterioration over time, essential for unraveling disease complexities. Nonetheless, analyzing these images is a labor-intensive task for specialists.

This research leverages data from the Cytology Laboratory of the Hematology service at 12 de Octubre Hospital in Madrid, Spain. The complexity observed in Bone Marrow Aspirate Smears (*BMAS*), which are rich in white blood cells at various stages of maturation, underscores the analytical challenges. Currently, the Differential Cell Count (*DCC*) technique is manually performed by many professionals, counting between hundreds of cells to diagnose blood diseases like leukemia [1]. This method is cumbersome, highlighting the necessity for an automated system that could be broadly adopted across health centers [11, 19, 22].

The objective of this work is to propose a computer vision system that employs deep learning to detect and classify instances in *BMAS*, aiming to streamline and enhance the accuracy of diagnosing diseases such as Acute Myeloid Leukemia (*AML*). This system utilizes an untapped dataset, ensuring it performs effectively in the typical conditions encountered in digital microscopy images. This study aims to answer the following research questions: (1) Can a deep learning model effectively automate the detection and classification of leukemia cells in bone marrow aspirate smears? (2) How does the proposed model compare to existing methods in terms of accuracy and precision?

This work is structured as follows. Section 2 reviews related works concerning the proposed method. Section 3 presents the methods and materials used, including the datasets and the proposed algorithm. Section 4 shows the results of multiple experiments performed. Finally, Sect. 5 outlines some conclusions and future work.

2 Related Works

Recent works in computer vision techniques applied to cell imaging have significantly expanded, showcasing an increasing interest [16]. Efforts to automate traditionally manual processes like the pap smear are evolving with unsupervised methods for effective segmentation and categorization of cervical cells [8], alongside innovative frameworks utilizing supervised machine learning for accurate identification of cell structures such as megakaryocytes [20]. Additional research has been directed towards enhancing the segmentation and feature extraction of white blood cells from blood smears, aiming to rectify limitations of previous works [2]. Moreover, the development of web-based systems facilitates the management of vast numbers of cells from digital pathology images, contributing to the standardization of clinical procedures such as the *DCC* [6].

In parallel, other studies have introduced fully automatic methods for identifying white blood cells in microscopic images, with some focusing on the use of deep learning for classifying acute leukemia [21]. Techniques that integrate both spectral and spatial features using algorithms like SVM-recursive feature elimination demonstrate increased effectiveness in cell identification [22]. Furthermore, diagnostic strategies for distinguishing between normal and myeloma cells in bone marrow smears are being refined, underscoring the potential for computer vision in medical diagnostics [18]. The exploration of feature space clustering and the comparison of color spaces highlight the ongoing enhancements in the

precision of cell segmentation from peripheral blood smear images [10]. These cumulative research efforts are instrumental in pushing the boundaries of digital pathology, facilitating quicker and more accurate medical diagnoses. Finally, [3] outlines a systematic and fully automatic process for identifying and classifying white blood cells using microscopic images, with Table 1 summarizing these methods for blood smear image segmentation.

Table 1. Comparison of methods with the proposed algorithm in [2].

Method	Image dataset	Accuracy	Size	References
Segmentation using threshold technique	Image captured from CDC-DPDx	98.4%	100	[4]
Segmentation using Otsu’s threshold algorithm	Public Image Dataset	93%	108	[11]
Segmentation by applying pre-processing at every stage	ALL-IDB	92%	108	[17]
Segmentation and classification using convolution neural network	ALL-IDB	96.42%	108	[21]
Segmentation using microscopic hyperspectral imaging technology means	Ruijin Hospital, Shanghai, China	92.9%	135	[22]
Segmentation using K-Clustering	Pathology department of Alzahra hospital, Isfahan, Iran	95.28%	50	[18]
Segmentation using scale space filtering and watershed clustering	–	98.9%	45	[10]
Segmentation using Zack’s threshold and distance conversion technique	Imam Reza (AS) hospital, Mashhad	93%	29	[3]

Unlike previous approaches that rely heavily on manual feature extraction, our method integrates both spectral and spatial features automatically using deep learning, which improves the segmentation accuracy in complex images. Additionally, our approach addresses the challenge of overlapping cells by incorporating a novel multi-class segmentation strategy, which is not present in other works.

3 Materials and Methods

3.1 Datasets

In the present study, two datasets have been used. In the following subsections, their distribution is described in detail and their limitations are discussed.

A Single-Cell Morphological Dataset of Leukocytes from AML Patients and Non-malignant Controls (AML—Cytomorphology-LMU)

According to [14] The Munich AML Morphology Dataset contains 18,365 (400×400 pixels) expert-labeled single-cell images taken from peripheral blood smears of 100 patients diagnosed with AML at Munich University Hospital between 2014 and 2017, as well as 100 patients without signs of hematological malignancy. At 100-fold optical magnification and oil immersion, images were captured using an M8 digital microscope / scanner (Precipoint GmbH, Freising, Germany). It was used to create all single-cell pictures in this dataset from peripheral blood smears at 100x magnification. The manufacturer claims a coverage of 14.14 pixels per micron. This database is available at TCIA [12].

Trained professionals categorized pathological and non-pathological leukocytes using a conventional morphological categorization system developed from clinical experience [14]. A subset of images was re-annotated up to two times to quantify inter- and intra-rater variability of examiners. The dataset was utilized to train a convolutional neural network for single-cell morphology classification, according to the scientists. There are 15 morphological classifications in this dataset, which may be divided into 3294 blasts (myeloblasts and monoblasts) and 15,071 nonblasts (the other 13 classes).

After processing the data, some images were blurred or had annotations that varied with each review by specialists. This could significantly affect the learning process of the algorithm. For this reason, it was decided to use a portion of the entire *AML* Dataset. This distribution is shown in Table 2. Some images are shown in Fig. 1 to illustrate how this dataset is given.

BMS—12 de Octubre Hospital Every day, the cytology laboratory at the 12 de Octubre Hospital’s hematological department conducts numerous *DCC* assays on peripheral blood smear and bone marrow smear slides. These data are gathered, analyzed, and kept as part of the diagnosis, as well as serving as a reference for future instances, as they offer valuable information to other health experts. The laboratory just purchased a set of cameras for its microscopes, allowing for uniform and high-quality digital picture recording. This enables the

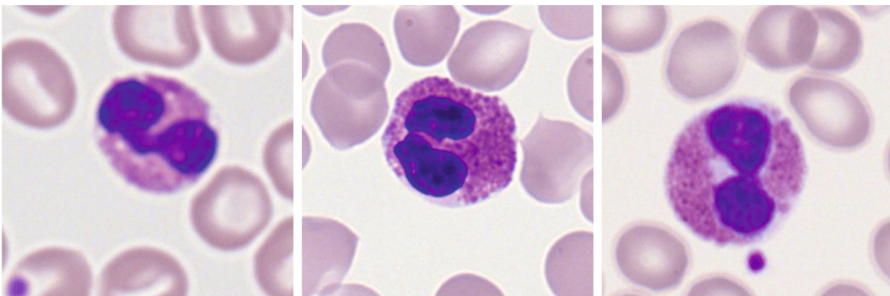


Fig. 1. Example of *AML* dataset, some images present a blur which must be corrected or eliminated in the training.

Table 2. Abbreviations of morphological classes used in folder structure and annotation file. All cells extracted from *AML* Dataset.

Name	Abbreviation	Amount
Basophil	BAS	69
Erythroblast	EBO	55
Eosinophil	EOS	420
Smudge cell	KSC	4
Lymphocyte (atypical)	LYA	8
Lymphocyte (typical)	LYT	3815
Metamyelocyte	MMZ	15
Monoblast	MOB	13
Monocyte	MON	1665
Myelocyte	MYB	32
Myeloblast	MYO	3150
Neutrophil (band)	NGB	101
Neutrophil (segmented)	NGS	1593
Promyelocyte (bilobed)	PMB	15
Promyelocyte	PMO	51
Unknown	UNC	52
Total		11,058

development of *BMAS* datasets, which may be very helpful in artificial intelligence applications; therefore, accomplishing work with this data is a significant step forward in this field.

The first images of the dataset were acquired in 2020 and correspond to more than 20 people with a range of blood cancer diseases at various stages, as well as healthy people. There are a total of 233 images in this collection, and the quality of the photographs is often poor. Data gathered for research activities should preferably be those that meet the optimal circumstances, with the goal of not jeopardizing the study. Images for medical diagnosis, on the other hand, are not taken with accuracy and are acceptable as long as the visualization is enough, since the focus in these instances is the patient’s health, and exposing the patient to additional samples to get better data may be deemed inappropriate. We did not include any images in this dataset that are challenging for the algorithm to learn. The laboratory collected images from patients in the following cases in terms of pathology: Normal bone marrow, Chronic lymphocytic leukemia, Non-Hodgkin lymphoma, Non-Hodgkin lymphoma (Burkitt type), Acute myeloid leukemia, Acute lymphocytic leukemia, and Multiple myeloma.

In May 2021, more *BMAS* smear images are collected and processed at the Hematology Service of the 12 October Hospital. This time, another 104 images were added to the database for this study. Of all images, 33 of them were pro-

cessed by specialists who, using a web tool explained in next section, marked each cell with its corresponding type to subsequently create the necessary masks for the algorithm to be used. The distribution of each cell type in this dataset is shown in Table 3. Figure 2 shows how this dataset is given.

Table 3. Abbreviations of morphological classes used in folder structure and annotation file. All cells extracted from 12 de Octubre Hospital 2021.

Name	Abbreviation	Amount
Eosinophil granulocyte	EOG	20
Myeloblast	MYO	21
Promyelocyte	PMO	12
Myelocyte	MYB	65
Metamyelocyte	MMZ	35
Neutrophil (band)	NGB	44
Neutrophil (segmented)	NGS	174
Promonocyte	PRM	1
Monocyte	MON	24
Basophilic Erythroblast	EBO	9
Polychromatic erythroblast	PLE	24
Orthochromatic erythroblast	ORE	88
Lymphoblast	LPB	31
Lymphocyte	LPC	80
Plasmatic	PLS	149
Total		777

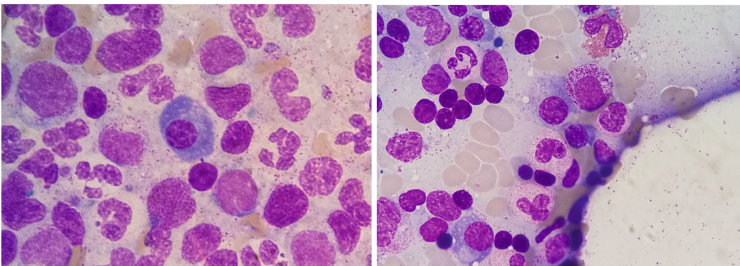


Fig. 2. Some images belonging to the 12 de Octubre Hospital 2021 dataset.

3.2 Annotation Tools

Once the datasets were correctly ordered, we proceed to make the respective annotations for each cell. In previous annotation process by the experts, a medical professional needs to mark the cells so that they can be correctly labeled later. With the intention of facilitating this process, a very intuitive tool is used. This tool allows you to mark each cell with a point and thus generate a file which later serves as an indicator for creating masks for each cell. Figure 3 (left) shows how these temporary images are marked, which show the number corresponding to each category and then based on that, draw the mask on the corresponding original image.

Once the annotations are clear, we proceed to create the masks that are necessary to train the algorithm. In this work **COCO-annotator** was used as an image annotation tool. COCO-annotator¹ is a web-based image annotation application that allows users to easily and easily label pictures to provide training data for image localization and object recognition. The generated file is downloaded. This contains the information needed for the training algorithms. The annotated datasets look as shown in Fig. 3 (right).

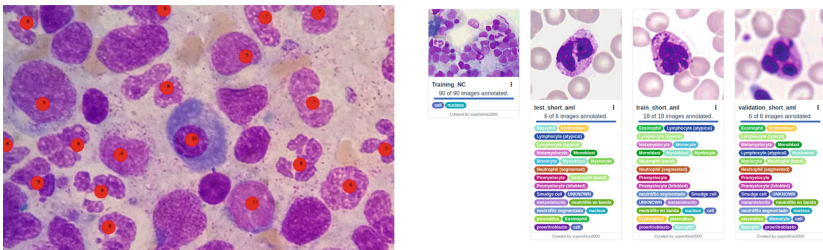


Fig. 3. **Left:** Visualization of the annotations previously made. **Right:** Examples of annotated datasets in COCO Annotator.

3.3 Evaluation Criteria

The Multi-Otsu thresholding method segments image pixels into distinct classes by intensity, using three thresholds to identify background, cell background, cytoplasm, and nucleus [7]. It removes components below the second threshold and uses morphological operations and smoothing for noise reduction. Following segmentation, classification and feature extraction processes are conducted. Classification uses a confusion matrix to evaluate true positives, true negatives, false positives, and false negatives, crucial for accuracy, precision, recall, and specificity metrics. Feature extraction focuses on cytomorphological characteristics like nucleus size, shape, area, perimeter, area-to-perimeter ratio, equivalent diameter, and nucleus-to-cytoplasm area ratio, with cell diameter calculated between the longest axis and its orthogonal counterpart.

¹ More information: <https://github.com/jsbroks/coco-annotator/>.

$$N : C = \frac{\text{Area of cytoplasm}}{\text{Area of cell}} \quad (1)$$

The proposed system generates multiple rectangular boxes and labels, indicating the object’s border, category, and location. The Intersection over Union (IoU) measures the accuracy of the projected border, while Mean Average Precision (mAP) assesses the accuracy of category labels. IoU quantifies the overlap between predicted and ground-truth boundaries, with a higher IoU indicating a better match. An IoU of 1 denotes complete overlap, while 0 indicates no overlap.

$$\text{IoU} = \frac{\text{Predicted boundary} \cap \text{Ground - truth boundary}}{\text{Predicted boundary} \cup \text{Ground - truth boundary}} \quad (2)$$

Mean Average Precision (mAP) represents the average precision across different recall levels, capturing the area under the Precision-Recall curve. This metric combines recall and precision to provide a comprehensive assessment of a model’s classification performance.

$$AP = \int_0^1 p(r)dr \quad (3)$$

Model predictions rely on confidence values from 0 to 1 for each data sample. According to [13], a sample is classified as positive if its confidence exceeds a predefined threshold; otherwise, it is negative. The choice of this threshold significantly impacts the balance between accuracy and recall (the proportion of actual positives correctly identified). In object detection, mAP and IoU are essential metrics for evaluating performance. A common IoU threshold is 0.5, where a predicted bounding box is correct if its IoU with the ground-truth box exceeds this value. Increasing predicted boxes generally enhances recall but may affect accuracy. mAP is computed by averaging precision across various recall levels and then taking the mean across all object categories, providing a comprehensive assessment of the model’s accuracy and error management.

3.4 Augmentation

Once the evaluation techniques have been defined, an important step in the classification phase is necessary. The problem is that the presence of different cell types is totally unbalanced. For example, in Fig. 4A, the distribution of cells is shown with the most abundant Lymphocytes with more than 3000 samples. While other cell types are minimal. Therefore, data enhancement is applied. This technique is used to augment the amount of data by adding slightly modified copies of existing data or newly created synthetic samples to existing data. Once the data is balanced, it is ready for the training phase. It is important to be clear that it is not necessary to have the same number of samples in each category, this can be seen in Fig. 4B.

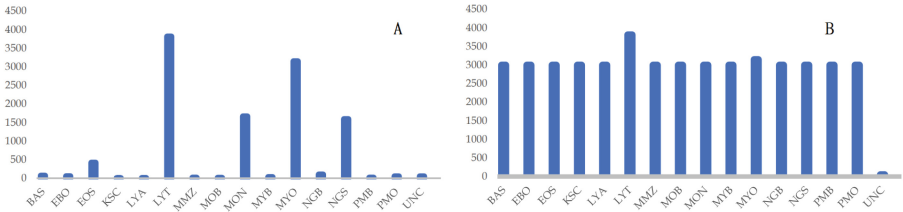


Fig. 4. Unbalanced distribution (A) and balanced distribution (B) of cells presented in the AML dataset (refer to Sect. 3.1).

3.5 YOLACT Algorithm

YOLACT training begins by displaying mAP values after each epoch. In this study, we used ResNet-50 as the backbone due to its 50-layer deep architecture, which is effective for detailed feature extraction, crucial for cell classification. The model was implemented and tested on a PyTorch platform using the code available on GitHub²

Previously, we used the Mask-RCNN algorithm [9] for cytoplasm and nucleus extraction, which is essential for obtaining relevant information. However, YOLACT [5] was chosen for this study because it allows instance segmentation through parallel tasks, generating prototype masks and predicting mask coefficients per instance. Additionally, YOLACT has a constant computational cost, unlike Mask-RCNN, where the cost increases with the number of instances.

4 Results

The AML and 12 de Octubre dataset are used in different task. The idea is to compare how this algorithm behaves in different situations, for example in the presence of many cells in a single image, absence of certain categories or datasets from different sources. The algorithm accepts the individual dimensions of each image which is not something that needs to be modified.

In the **Experiment A**, the subdataset is composed of 18 training images and 6 validation images. This dataset contains 3 classes and the data is from *AML* Dataset. The results in Fig. 5 shows that the trained model can segment cells in pixel level. This is important because we are getting closer and closer to the next work objectives.

In the **Experiment B**, the subdataset was composed of 90 training images and 11 validation images. The images belong to the Hospital 12 de Octubre 2020-2021. This subdataset contains 2 classes as follow, and every single image contains several of them. After one hour of training, the results are processed to obtain early detections of cells and nuclei. The results show that the algorithm detects some instances but not most. This represents a better search for parameters and possible complications. For example in Fig. 6A, the training time was

² More information: <https://github.com/dbolya/yolact>.

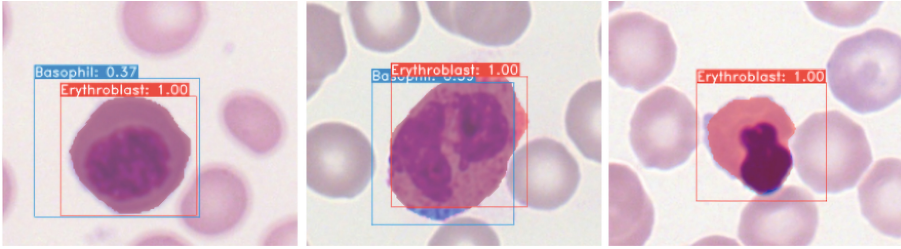


Fig. 5. Results of different stages of segmentation from the AML dataset. The images are three examples. The images depict how the model isolates individual cells for further classification.

20 minutes. It can be seen that the algorithm generates a mask larger than normal (bluish color) and fully expanded over the image. While Fig. 6B does not have a mask outside the normal size, but the cells and nuclei that it detects are not complete since they present a *cell score* between 0.50 to 0.70 in some cases. Finally, Fig. 6C shows a better result but as there are few images there are no significant improvements.

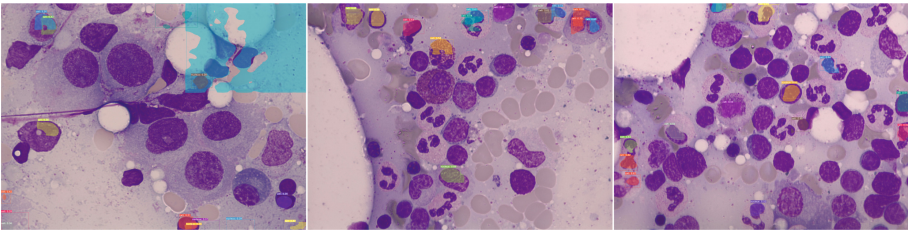


Fig. 6. Results in Experiment B. **A:** Sample 1. A large mask created by lack of training is shown. **B:** Sample 2. The algorithm marks some cells with errors in their masks. **C:** Sample 3. It shows a better result in the masks but does not detect all the cells.

In the **Experiment C**, the subdataset is composed of 93 training images and 11 validation (23 different classes). Images are from AML Dataset. With this experiment, the algorithm is allowed to run a larger number of epochs and these weights are stored. In these results (Fig. 7) is observed a great ability to detect the cells and categorize them. In this experiment, a confusion matrix is not developed as it is not the objective of this dataset (although the results evaluated were all well classified).

Nucleus and Cells A subdataset with a combination of *AML* dataset, 12 de Octubre Hospital 2020 and 2021, is generated to train the model. The training parameters were 10 epochs with 18 steps per epoch. The training time was 120 minutes. The model classified all images in the training set correctly, with 96.2%

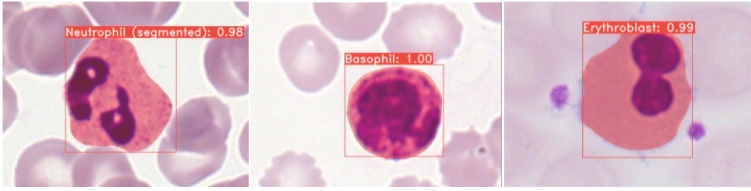


Fig. 7. Results in Experiment C. These results show a good performance of the algorithm to detect and classify the cells belonging to the *AML* Dataset.

accuracy on testing set. Precision, recall, and specificity values for the testing set were above 90%. The model achieved a high accuracy in the binary classification task, correctly identifying 89 nucleus instances and 92 cell instances, with only 4 and 3 misclassifications, respectively.

Figure 8A and B are shown. These show an approximate calculation of the diameters of each detection. This is because there are no exact measurements from the dataset. In addition, by combining different training samples, it is not possible to make a more accurate estimate. Finally, Figure 9 shows the algorithm output for one of the test images. In this image, not all instances were detected, however a large percentage were detected thanks to the data augmentation techniques developed by the algorithm.

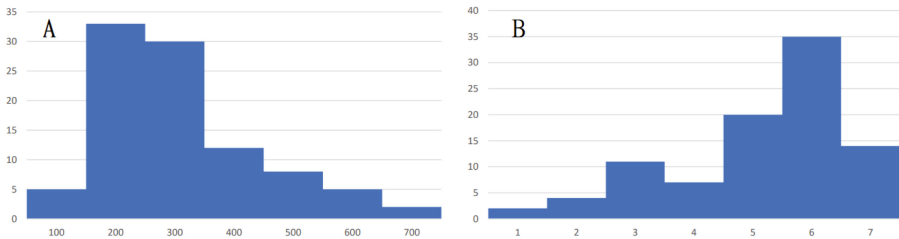


Fig. 8. Frequencies of cytoplasm detected (A) and detected nuclei (B).

Cell Classification This part is one of the most important because the final approach of the algorithm is tested. In this part, the training dataset combines all the other datasets. For this phase, as mentioned at the beginning, it is a process that could be done manually and that in a certain way these outputs could be solved by just making a script to store the predicted and real categories.

The system uses the same training parameters to train the algorithm, except for the training time, which can vary according to other variables and dataset sizes. Figure 10 shows a confusion matrix with the data resulting from the classification of each of the 20 cell types. Table 4 shows the calculated metrics of the classification. In Fig. 11, the size distributions of the detected cells are shown.

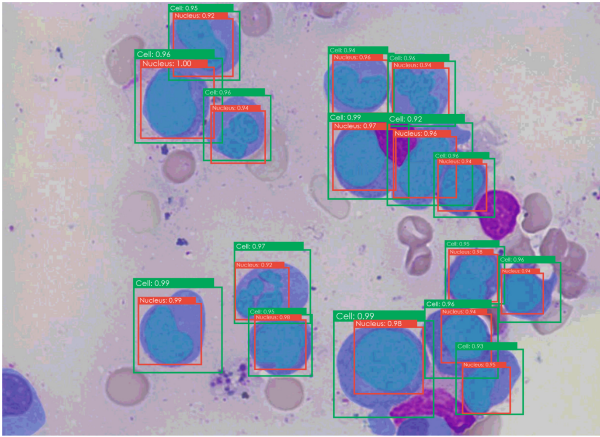


Fig. 9. One of the testing images from the dataset 12 de Octubre 2020.

Table 4. Performance metrics of the model for classification between 20 types of cells.

Set	Accuracy	Precision	Recall	Specificity
Training set	98%	97%	98%	97%
Testing set	98%	96%	97%	96%

Figure 12 shows one of the images that were used for the testing process. The scores show a great performance of the algorithm.

5 Conclusions

To overcome the limitations of existing manual methods in *BMAS* analysis, a novel computer vision approach is proposed. This system performs automatic detection and classification of cells present in this type of sample. The performance of the proposed system is very good, as the results yield more than 90% accuracy using limited data and computational capacity.

In this study, we work with two datasets. The first one is AML - Cyto-morphology_LMU. On the other hand, we have two sub-datasets collected by specialists of the Hospital 12 de Octubre in Madrid. The first subdataset was collected in 2020 and has approximately 233 images. While the second part was collected in 2021 and has more than 30 images that were annotated and are filled with cells from a total of 20 different categories.

The proposed system represents a significant step forward in assisting clinical specialists. For instance, it could be integrated into hospital workflows to automatically pre-screen bone marrow aspirate smears, providing preliminary classifications that hematologists can review and confirm. However, to make this system a reality in clinical settings, further steps such as regulatory approval,

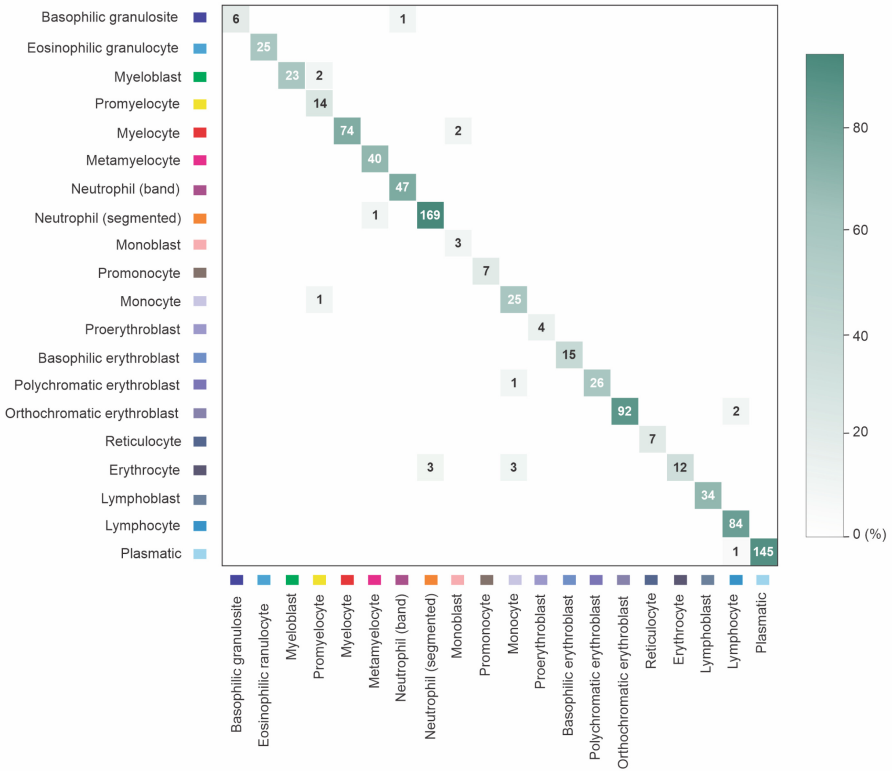


Fig. 10. Confusion matrix with 20 different cell categories.

integration with existing hospital information systems, and extensive validation studies across diverse patient populations are necessary. Our ResNet-50 outperformed simpler models like ResNet-18 due to its deeper architecture, which captures more detailed features crucial for cell type distinction. However, this

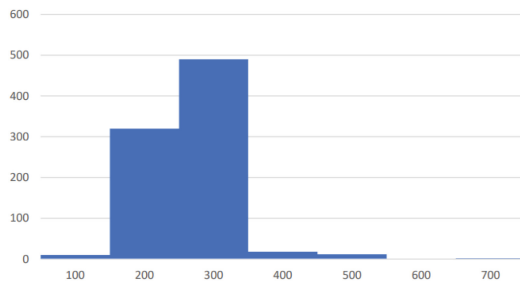


Fig. 11. Histogram of frequencies in the detected cytoplasm of 20 types cell types.

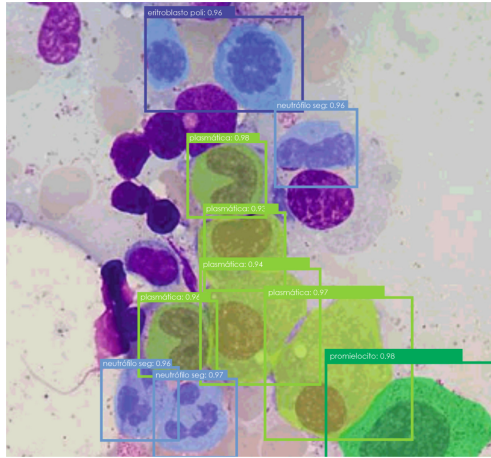


Fig. 12. One of the images used in testing from the dataset 12 de Octubre 2021.

increased complexity also raises computational costs, suggesting a need to balance model depth with efficiency for clinical use.

This study has some limitations that should be addressed in future research. One significant limitation is the small size of the dataset obtained from Hospital 12 de Octubre, which may affect the generalizability of the results. Additionally, potential biases inherent in this dataset could influence the model's performance. These limitations suggest that larger and more diverse datasets are needed to improve the accuracy and reliability of the proposed system. Future work should also consider methods to mitigate any biases that may arise from the data used. Furthermore, more information from the Hospital 12 de Octubre images is needed to improve accuracy in detection and classification. Additionally, ranking the importance of cell features could enhance the system's ability to distinguish between similar cells and focus on particular diseases. The proposed approach serves as an effective tool for specialists by reducing data analysis time and generating statistics. Further development and collaboration with hospitals could lead to wider adoption and advancement of such systems in cytology.






Acknowledgements. This work is part of the R+D+i projects PID2019-108915RB-I00 and PID2022-140907OB-I00 as well as by the grant PRE2020-094056 funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU. It has also been partially supported by Junta de Comunidades de Castilla-La Mancha/ERDF (SBPLY/21/180501/000030) and by the University of Castilla-La Mancha (2022-GRIN-34436).

References

1. Abdulrahman, A., et al.: Is a 500-cell count necessary for bone marrow differentials? *Am. J. Clin. Pathol.* **150**, 84–91 (2018)
2. Acharya, V., Kumar, P.: Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms. *Med. Biol. Eng.* (2019)
3. Alreza, Z.K.K., Karimian, A.: Design a new algorithm to count white blood cells for classification leukemic blood image. In: 2016 6th ICCKE (2016)
4. BelekarS., J., Chougule.S, R.: WBC segmentation using morphological operation and SMNT operator. *Int. J. Innov. Res. CCE* **3** (2015)
5. Bolya, D., et al.: YOLACT: real-time instance segmentation. *CoRR abs/1904.02689* (2019)
6. Chandradevan, R., et al.: Machine-based detection and classification for bone marrow aspirate differential counts. *Lab. investig.* **100** (2019)
7. Feng, Y., et al.: A multi-scale 3d OTSU thresholding algorithm for medical image segmentation. *Dig. Sig. Process.* **60** (2016)
8. Gençtav, A., Aksoy, S., Önder, S.: Unsupervised segmentation and classification of cervical cell images. *Pattern Recogn.* **45**(12), 4151–4168 (2012)
9. He, K., et al.: Mask R-CNN. In: 2017 IEEE ICCV, pp. 2980–2988 (2017)
10. Jiang, K., et al.: A Novel White Blood Cell Segmentation Scheme Using Scale-Space Filtering and Watershed Clustering, vol. 5, pp. 2820–2825 (2003)
11. Joshi, M.D., Karode, A.H., Suralkar, P.S.R.: White blood cells segmentation and classification to detect acute leukemia ms (2013)
12. K, C., et al.: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository (2013)
13. Liang, X.: Chapter 6—case studies. In: Liang, X. (ed.) *Ascend AI Processor Architecture and Programming*, pp. 229–288. Elsevier (2020)
14. Matek, C., et al.: A Single-cell Morphological Dataset of Leukocytes from AML Patients and Non-Malignant Controls [Data set] (2019)
15. National Cancer Institute: Leukemia (2024). <https://www.cancer.gov/types/leukemia>. Accessed 19 Aug 2024
16. Puchaicela, M., et al.: Deep learning for glaucoma detection: R-CNN Resnet-50 and image segmentation. *J. Adv. Inf. Technol.* **14** (2023)
17. Putzu, L., Ruberto, C.D.: White blood cells identification and counting from microscopic blood images (2013)
18. Saeedizadeh, Z., et al.: Automatic recognition of myeloma cells in microscopic images using bottleneck algorithm. *J. Microsc.* **261**(1), 46–56 (2016)
19. Saltos, K., et al.: Detecting Parkinson’s disease with convolutional neural networks: voice analysis and deep learning. In: *Information and Communication* (2023)
20. Song, T.H., et al.: A Circumscribing Active Contour Model for Delineation of Nuclei and Membranes of Megakaryocytes, vol. 9420 (2015)
21. Ttp, T., et al.: Acute leukemia classification using convolution neural network in clinical decision support system (2017)
22. Wang, Q., et al.: Spectral-spatial feature-based neural network method for acute lymphoblastic leukemia cell identification. *Biomed. Opt. Exp.* (2017)



Breast Thermographic Image Augmentation Using Generative Adversarial Networks (GANS)

Ramiro Israel Vivanco Gualán^{1,2} , Yuliana del Cisne Jiménez Gaona^{2,3,4} ,
Darwin Patricio Castillo Malla^{2,3,4} , María José Rodríguez-Alvarez³ ,
and Vasudevan Lakshminarayanan⁴ 

- ¹ Departamento de Ciencias de La Computación, Universidad Técnica Particular de Loja, San Cayetano Alto S/N CP1101608, Loja, Ecuador
rivivanco@utpl.edu.ec
- ² Departamento de Química y Ciencias Exactas, Universidad Técnica Particular de Loja, San Cayetano Alto S/N CP1101608, Loja, Ecuador
{yjdjimenez, dpcastillo}@utpl.edu.ec
- ³ Instituto de Instrumentación Para La Imagen Molecular I3M, Universitat Politècnica de Valencia, 46022 Valencia, Spain
mjrodri@i3m.upv.es
- ⁴ Theoretical and Experimental Epistemology Lab, School of Optometry and Vision Science, University of Waterloo, Waterloo, ON N2L3G1, Canada
vengumo@gmail.com

Abstract. Breast thermography captures infrared radiation images to monitor skin surface temperature changes non-invasively. This data, when combined with artificial intelligence, facilitates early breast cancer diagnosis and detection. However, training deep learning algorithms such as convolutional neural networks is challenging due to the limited number of images. The primary objective of this study is to create a set of synthetic breast thermographic images using segmentation and data augmentation techniques. In this work, we propose 1) Using public breast thermography databases, 2) Segmenting the region of interest with the U-Net network, 3) Increasing the variety of thermographic images using the SNGAN model, and 4) Evaluating the performance and accuracy of the previous algorithms with statistical metrics. The results indicate that the U-Net achieved an IoU of 0.96 and a Dice coefficient of 0.97. The SNGAN network generated 2000 synthetic images, reflected in a KID value of 4.54. In conclusion, U-Net is highly effective for segmenting regions of interest in thermographic images, and SNGAN shows promising results in synthetic image generation.

Keywords: Data augmentation · Segmentation · SNGAN · Thermography breast images · U-Net

1 Introduction

Technological and scientific advances in biomedicine and radiomics have led to the development of innovative tools for the early detection of diseases. One emerging tool in this field is breast thermography, a non-invasive method that has gained relevance in

breast cancer detection [1]. Despite its potential, thermography remains a complementary technique to others, such as ultrasound or mammography, which are helps in breast cancer screening and diagnosis.

Breast cancer presents specific disease characteristics, such as angiogenesis, which can be detected through thermography due to the heat generated in the affected area [2]. Thermography is based on infrared radiation, a type of radiation that indicates an object's temperature and can be used in pathology detection [3].

Research on thermal flow and transfer in biological organisms, as demonstrated by Pennes (1948), creates a foundation for understanding perceived thermal variations. The integration of artificial intelligence, specifically through the use of deep learning algorithms and neural networks, proposes the capability to significantly improve identification and precision in the early stages of breast cancer diagnosis [4–6].

The acquisition of a larger dataset for validation and training poses a challenge. Therefore, numerous data augmentation techniques have emerged, including color space augmentations, geometric transformations, image mixing, or kernel filters [7]. However, deep learning techniques offer new alternatives and synthetic data more realistic based on the distribution for synthetic data generation, such as autoencoders [8], generative adversarial networks (GANs) [9], and neural style transfer [10]. These techniques use the distribution of images to generate high-quality synthetic images. GANs have been highly successful in synthesizing medical images. Artificially generated images by GANs [11], such as Cycle GAN [12], Conditional GAN [13], WGAN [14], Vanilla GAN [9], Style GAN, and Progressive GAN [15], are very realistic.

GANs hold particular importance [7] because they have proven effective in generating high-quality images [12, 13, 16], particularly variants like SNGAN. This technique has gained popularity for generating high-quality synthetic data.

2 Methodology

The methodological design is based on five phases: (i) Data Collection, (ii) Data Preparation, (iii) Segmentation, (iv) Data Augmentation, and (v) Statistical Evaluation [17–21], as shown in Fig. 1.

2.1 Collection of Breast Thermographic Images

From public databases, 615 breast thermographic images were collected, categorized as malignant and benign classes, and non defined, as shown in Table 1. These images were taken from different views and angles, and 209 frontal breast images were selected for the research. These images had variations in size, resolution, color, noise, and the most contains artifacts as we can observed in Fig. 2. Therefore, the images had to undergo preliminary preparation before their use in the segmentation algorithm.

2.2 Images Preprocessing

This phase involved subjecting the 209 frontal breast images to a normalization and cleaning process to ensure their homogeneity. Firstly, the images were resized to a standard resolution of 450 x 300 pixels, as it was the common area of interest in most images.

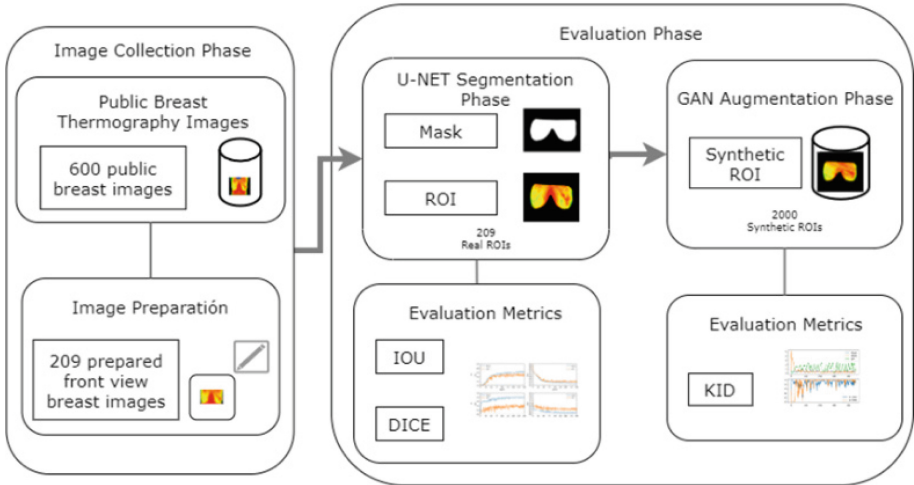


Fig. 1. Proposed methodology.

Table 1. Public repositories of breast thermographic images

Datasets	Link	Benign	Malignant	Total
Visual Lab, breast thermal image - Brazil	https://bit.ly/3Zhahen	186	150	336
Irthermo Benchmark Dataset Collection - University of Rajasthan Iran	https://bit.ly/3Rz83FF	50	32	82
Irthermo Benchmark Dataset Collection - University of Rajasthan Iran	https://bit.ly/3Rz83FF	50	32	82

Then, artifacts were removed, and segmentation masks based on Thermal Anthropometry were created using tools like Paint.net and Image J to delineate and extract areas of interest.

Subsequently, the information was organized into a CSV file named “breast_data.csv,” which indexes details such as image name, classification, and type of view. As previously mentioned, only frontal view images (see Fig. 2) were used in this work, although other views were preserved for future research. The entire dataset was stored in a.zip file and uploaded to a Google Drive account, accessible via the following link: <https://bit.ly/3Pib8ah>.

2.3 Segmentation with U-Net

The segmentation of breast thermographic images (see Fig. 3) was carried out using the U-Net network architecture on a Google Colab virtual GPU. The images and masks

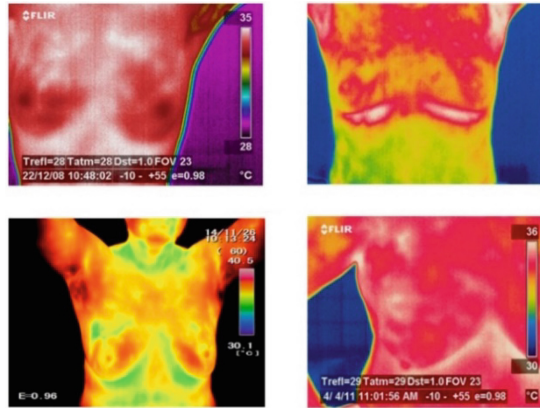


Fig. 2. Breast thermographic images. (a) Camera information on the image (b) Low image resolution (c) Different color gamut (d) Breasts outside the image area.

were normalized to a size of 512×512 pixels, considering that U-Net is more efficient with image sizes that are multiples of 2. The masks were binarized so that the pixels had values between 0 and 1. The U-Net network architecture, designed to work with 512×512 pixel images and three-color channels, processes and adjusts approximately 31 million parameters during its training.

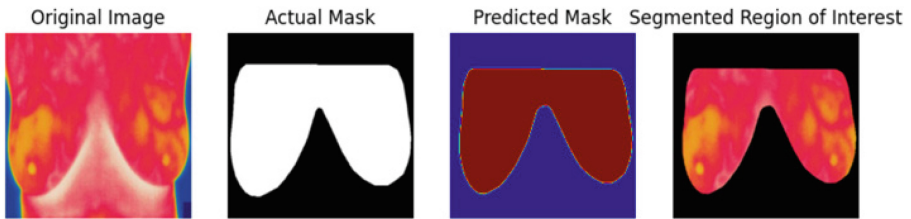


Fig. 3. Segmentation of the region of interest. (a) Original image (b) Extraction mask (c) Prediction mask, and (d) Segmented region of interest.

During training, the dataset was divided into three subsets with the following distribution: training (80%), validation (10%), and testing (10%). The performance of the U-Net network was configured using various hyperparameters. Table 2 details these hyperparameters, including a total of 30 epochs, a batch size of 4, and the use of the ‘rmsprop’ optimizer.

2.4 Image Augmentation with SNGAN

During the image augmentation phase, the Spectral Normalization GAN (SNGAN) network (see Fig. 4) used the previously segmented real breast regions of interest as

Table 2. U-Net hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Number of epochs	30	Training set	80%
Filter size	(3, 3)	Validation and test set	20%
Batch size	4	Output activation layer	‘sigmoid’
Pooling size	(2, 2)	Convolution activation layer	‘relu’
Optimizer	‘rmsprop’	Input size	(512, 512, 3)

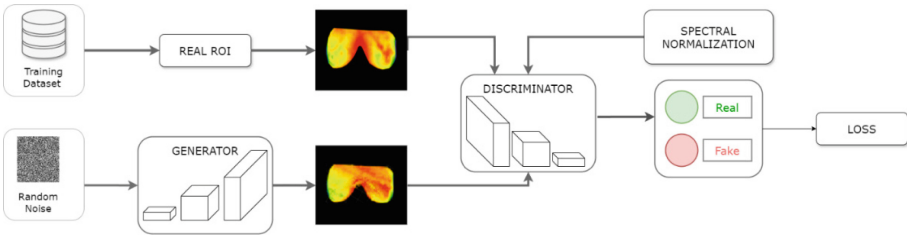


Fig. 4. SNGAN neural network architecture

input for the network’s discriminator and using the spectral normalization technique; it provided feedback to the generator to produce images increasingly like the original set.

Similar to the segmentation phase, a Google Colab Pro account was used to train the generator network and leverage the GPU power in order to process and convert images into tensors, which were optimized for high-performance calculations.

Table 3 details the general configuration of the training and generation process. A single V100 GPU was used for execution, and images of size 512×512 with three channels were processed. The training was performed in batches of 4 images, considering two distinct classes of images, benign and malignant. The latent space for image generation has a dimension of 10 and was implemented with the Adam optimizer for 150 epochs during training, with a learning rate of 0.00015. Additionally, a seed of 999 was set to generate 1,000 images by implementing spectral normalization with an epsilon of 1×10^{-12} .

For the generator parameters shown in Table 4, a ReLU activation function with a latent noise dimension of 10 was used; for backpropagation, the BCE with LogitsLoss loss function was employed. The generator model consists of 5 layers, starting with 1024 filters in the first layer and using a 4×4 kernel size. On the other hand, the discriminator employs the LeakyReLU activation function and performs five iterations. Like the generator, the discriminator used the BCE with LogitsLoss loss function. Its architecture consists of 8 layers, starting with 64 filters and a 4×4 kernel size.

Table 3. General SNGAN Hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Execution device	V100	Number of epochs	150
Batch size	4	Learning rate	0.00015
Number of image classes	2	Optimizer	Adam
Image size	512	Seed	999
Image channels	3	Number of images to generate	1000
Latent space dimension	10	Spectral normalization epsilon	1e-12

Table 4. SNGAN generator and discriminator hyperparameters

Generator hyperparameters	Value	Discriminator hyperparameters	Value
Activation function	ReLU	Activation function	LeakyReLU
Latent noise dimension	10	Discriminator iterations	5
Loss function	BCE with LogitsLoss	Loss function	BCE with LogitsLoss
Number of layers	5	Number of layers	8
Filters in the first layer	1024	Number of filters	64
Kernel size	4×4	Kernel size	4×4

2.5 U-Net Performance Evaluation

During the training of the U-Net network, two metrics were incorporated: IoU and the Dice coefficient. The IoU metric (Eq. 1), calculated with the mean_iou function, evaluates the overlap between the prediction and the actual mask, with its value ranging from 0 (no match) to 1 (perfect match).

$$IoU = (objective \cap prediction) / (objective \cup prediction) \quad (1)$$

The Dice coefficient (Eq. 2), implemented through dice_coef(), further weights the intersection, highlighting the prediction's accuracy.

$$DSC = 2|objective \cap prediction| / (|objective| + |prediction|) \quad (2)$$

2.6 SNGAN Performance Evaluation

The Kernel Inception Distance (KID) metric was used to evaluate the quality and diversity of generated images in relation to real images (Eq. 3). This metric employs the pre-trained InceptionV3 network for feature extraction.

During the training process, at the beginning of each epoch, the KID value was updated and recorded to monitor the quality of the generated images. Once each epoch was completed, this metric is reset for the next iteration. After completing the training, a graph is presented showing the evolution of the KID value over the epochs.

$$KID(p, q) = (\Phi(p) - \Phi(q))^2 \quad (3)$$

In (Eq. 3), $\Phi(p)$ and $\Phi(q)$ represent the Inception representations of the real and generated samples, respectively.

3 Results

The training, validation, and testing of the models for the segmentation and augmentation phases using their evaluation metrics presented the following results.

3.1 RoIs Extraction with U-Net

Throughout the training process, a file named “UNET.h5” was generated, containing the resulting weights. This file was used to load the weights into the model and perform the prediction with the prepared original image set. The images were then concatenated with the obtained predictions. The result, shown in Fig. 5, represents the set of synthetic regions of interest for both benign and malignant images.

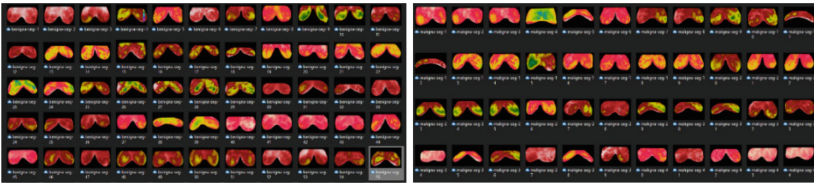


Fig. 5. Segmented images, (a) benign and (b) malignant

3.2 Training with U-Net

The loss revealed a significant difference between training and validation, suggesting effective generalization of the model, as seen in Fig. 6 a-b. The hyperparameters were adjusted in multiple iterations, focusing mainly on the number of batches and epochs to optimize training stability and accuracy. The hyperparameters were adjusted in multiple iterations, focusing mainly on the number of batches and epochs to optimize training stability and accuracy. This is consistent with other studies, such as Li et al., who also used U-Net for image segmentation in mammography, achieving an IoU of 0.853 and a Dice coefficient of 0.906. Similarly, Guo et al. reported an IoU of 0.912 and a Dice coefficient of 0.953 in their expanded U-Net model for breast ultrasound image segmentation. In comparison, our model achieved higher performance with a Mean IoU of 0.96 and a

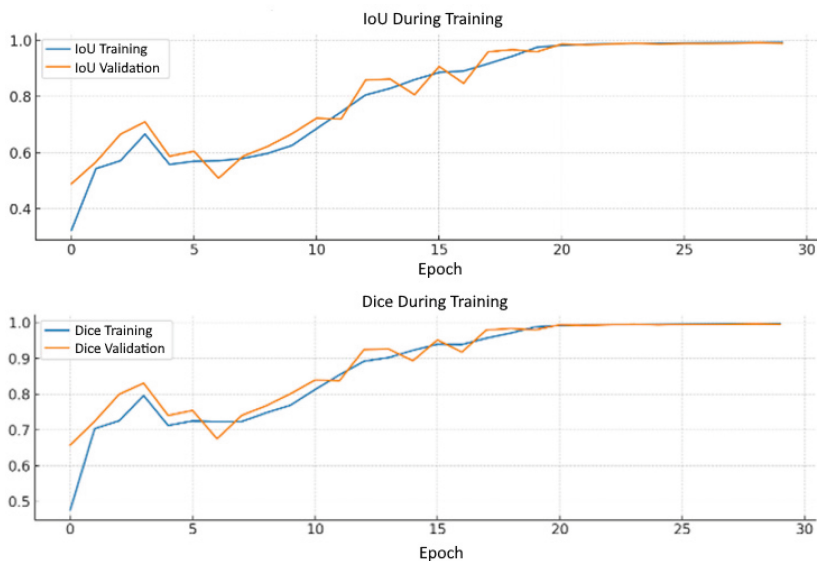


Fig. 6. Evaluation metrics of the U-Net algorithm during training (a) IoU and (b) Dice

Dice coefficient of 0.97, indicating the robustness of U-Net for thermographic breast image segmentation. The data in our study stabilized upon reaching epoch number 30, which is an adequate value to avoid overfitting.

Table 5 summarizes the mean values of the metrics, highlighting a solid overall performance of the model. High precision and specificity were observed, although sensitivity showed room for improvement. These metrics, especially the Mean IoU and the Dice Coefficient, reflect the model's ability to accurately identify the RoIs in the images. The set similarity and Jaccard coefficient metrics, in turn, underline robust concordance between the predictions and the true values.

3.3 Synthetic Thermographic Images

During the training of the SNGAN network, the generator aimed to produce images that challenged the discriminator's ability to distinguish them from real ones, and the discriminator evaluated them. Once the training was completed, the image quality was evaluated using the KID metric, and once the image quality was verified, the images were stored, resulting in approximately 2000 images. Several samples can be seen in Fig. 7.

The quality of the generated synthetic images was evaluated using the Kernel Inception Distance (KID) metric, which measures the similarity between real and generated images by extracting features through the InceptionV3 network. During each epoch of training, KID values were calculated by directly comparing the generated images with a set of real images. A lower KID value, recorded in each iteration, indicates that the synthetic images are increasingly resembling the real ones, reflecting an improvement in generation quality.

Table 5. Average values of established metrics

Metric	Mean value
Loss	42.23
Mean IoU	0.96
Dice coefficient	0.97
Validation loss	0.09
Validation mean IoU	0.82
Validation dice coefficient	0.89
Accuracy	0.82
Precision	0.97
Sensitivity	0.66
Specificity	1

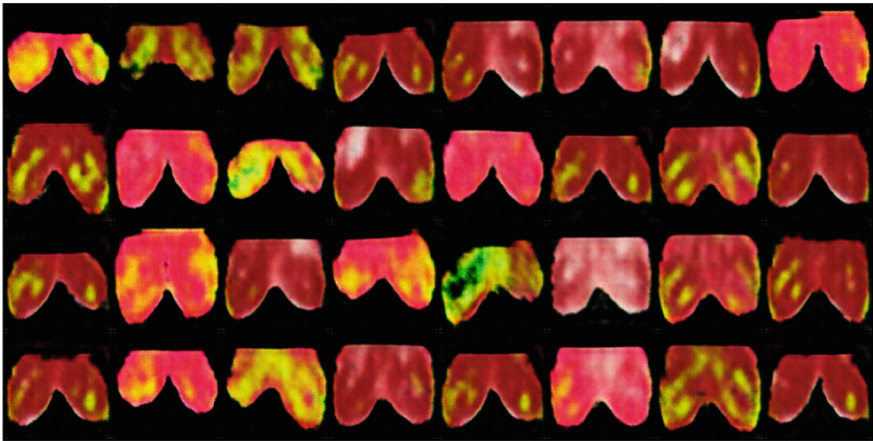


Fig. 7. Batch of synthetic images from segmented ROIs

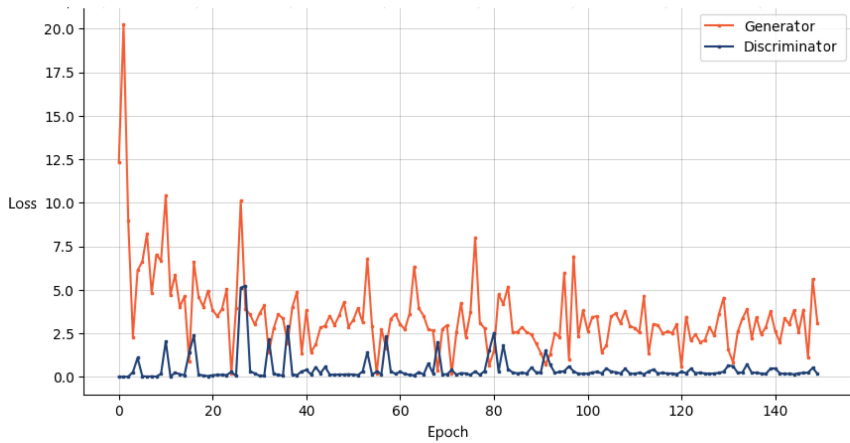
Generator vs Discriminator with SNGAN.

Table 6 reflects the results obtained in the augmentation phase, where the discriminator and generator were evaluated in pairs with a mean value of 0.32 and 3.9 respectively. The similarity of real vs. fake images was evaluated with a mean of 0.91 and 0.22 respectively, and finally, the image quality was evaluated using the KID metric with a mean value of 4.54. The resulting training graphs are shown below.

In Fig. 8, we observe that the discriminator and generator values exhibit diverse behaviors throughout the algorithm’s iterations. The graph indicates that the discriminator loss was optimal, beginning to stabilize after epoch 40. In contrast, the generator showed more unstable execution, though it began to normalize after epoch 60.

Table 6. Average values of SNGAN evaluation metrics

Metric	Mean value
Discriminator	0.326505
Generator	3.963454
Real	0.914982
Fake	0.222153
KID	4.547285

**Fig. 8.** Generator vs discriminator

3.4 KID Metric

Figure 9 shows the training results implementing the KID metric using the KID Metric class, evaluating the quality and diversity of synthetic images compared to real images using the Kernel Inception Distance (KID) metric. KID measures the Maximum Mean Discrepancy Squared (MMD) between the Inception representations of real and generated samples using a polynomial kernel [22]. This metric was used because the evaluation requires fewer samples since we do not need to adjust the square covariance matrix.

The Kernel Inception Distance (KID) metric was used to evaluate the quality and diversity of generated images in relation to real images (Eq. 3), employing the pre-trained InceptionV3 network for feature extraction. During the training process, at the beginning of each epoch, the KID value was updated and recorded to monitor the quality of the generated images. Studies such as Heusel et al. [18] have demonstrated that KID is particularly effective in evaluating the quality of GAN-generated images, offering advantages over the Frechet Inception Distance (FID) due to its lower bias with smaller sample sizes. In our study, the KID metric indicated a growing similarity between the generated and real images, which is consistent with findings in other research that applied KID to assess image quality in GAN-generated medical images. After completing the

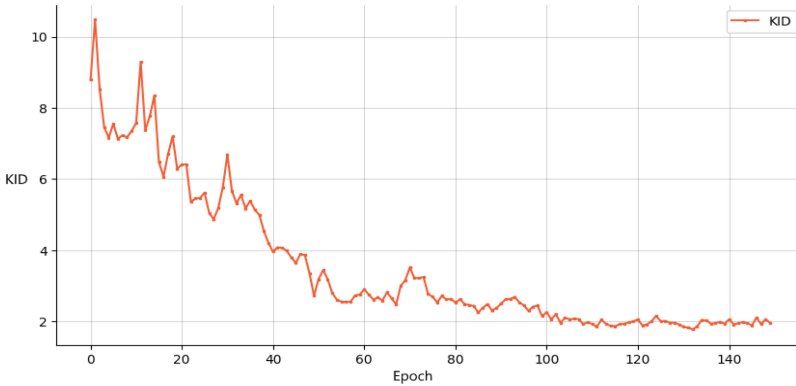


Fig. 9. KID evolution during training

training, a graph is presented showing the evolution of the KID value over the epochs, similar to what is reported in Betzalel et al. [22], where the KID metric was utilized to track the improvement in image quality during GAN training.

4 Discussion

4.1 Segmentation Phase with U-Net

The U-Net network demonstrated high accuracy in segmenting thermographic breast images, achieving an IoU value of 0.96 and a Dice coefficient of 0.97. These results show that U-Net was able to overlay the original image with the segmentation mask and accurately select only the region of interest to segment. The resulting values can be compared with those from other previously reported studies., such as Li et al. [23] who focused on mammography-based image segmentation using U-Net, achieving an IoU of 0.853 and a Dice coefficient of 0.906, or Guo et al. [24] where they applied an expanded U-Net network for breast ultrasound image segmentation, obtaining an IoU of 0.912 and a Dice coefficient of 0.953. Compared with these studies, it can be observed a significant improvement in the performance that the model has achieved over previous methods. This advance reinforces the utility of the U-Net network applied to the segmentation of thermographic images compared to results applied to other types of breast cancer-related images.

4.2 Image Augmentation Phase with SNGAN

The SNGAN model was able to generate a set of both benign and malignant breast thermographic images, achieving a KID value of 4.547285, which, although indicating high similarity between the original and synthetic images, still leaves room for improvement. When compared to previous research, such as Yusoff et al. [25], who used StyleGAN for generating histopathological images and achieved a KID value of 1.0, our results suggest that further optimization of SNGAN could enhance image quality. Additionally, studies

like those by Gupta et al. [18] using SNGAN for medical image synthesis also highlight the importance of a larger and more diverse training set, which could help reduce the KID value further. Despite this, the generated images in our study show significant potential for augmenting datasets in scenarios where larger sets are required for training neural networks on breast thermographic images.

4.3 Study Limitations and Future Work

A limitation of this study was the scarce availability of public breast thermography images, which limits the training of neural networks for generation and classification, as these algorithms require large amounts of data. To improve this, images generated by the SNGAN network could be used to refine segmentation and classification algorithms, thus increasing their effectiveness. It would be beneficial to explore other GAN architectures and optimize the hyperparameters of the SNGAN to enhance the quality of the images. According to Alfayez et al. [26], it is recommended to use the expanded dataset from this study to apply Support Vector Machines (SVM) for breast cancer detection.

Another potential limitation in this study may be related to the data sources used, which come from patients in specific countries such as Brazil, Iran, and India. As shown in the Table 1, there is a significant variation in the size of data from each region, with 336 images from Brazil, 197 from India, and only 82 from Iran. This disparity could introduce bias into the model, as certain population groups are much more represented than others, potentially limiting the algorithm's accuracy and generalization for populations from other countries not represented in these data.

In future work, it will be essential to include more diverse and representative datasets from a broader range of populations to assess whether the current model exhibits biases in its performance. The inclusion of new data sources would help generate synthetic images that better reflect global diversity, thereby improving the model's generalization and applicability in a wider clinical context.

5 Conclusion

This study explored the application of deep learning techniques to thermographic breast imaging, a non-invasive cancer detection method that benefits significantly from advancements in artificial intelligence. The challenges primarily arose from restricted data access due to patient privacy and protection. Our approach focused on segmentation and synthetic generation of regions of interest (ROIs) in thermographic breast images using the U-Net and SNGAN networks, respectively.

The U-Net network was highly effective in segmenting ROIs, demonstrating strong model performance through key metrics: an IoU coefficient of 0.96 and a Dice coefficient of 0.97. These metrics reflect the network's precision in segmenting areas of interest. The average accuracy achieved was 0.97, underscoring the model's ability to predict ROIs correctly. Although the sensitivity was slightly lower at 0.66, it is still considered acceptable in this context. The exceptionally high specificity of 0.99 indicates that U-Net effectively identified non-relevant areas.

In contrast, the SNGAN network, trained to generate new images from the ROIs segmented by U-Net, showed promising potential to augment the dataset of breast thermographic images. The consistent reduction in the KID metric during training indicates a growing similarity between the generated and original images, suggesting that SNGAN could effectively mitigate data access limitations.

Future studies could employ the generated images to bolster segmentation and classification algorithms, as well as explore and optimize other GAN architectures. The utilization of cloud platforms like Google Colab Pro for efficient processing is recommended, along with retraining the U-Net network with data generated by SNGAN and exploring other approaches like the combination of the Optimal Weight Neural Network (OWNN) with the Convolutional Neural Network (CNN). Additionally, it is recommended to apply Support Vector Machines (SVM) in breast cancer detection with the expanded dataset [26, 27].






References

1. P. Zamorano, "Análisis y Caracterización de Imágenes Termográficas de Mamas," Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, 2019. Accessed: Nov. 19, 2022. [Online]. Available: <https://inaoe.repositorioinstitucional.mx/jspui/handle/1009/1845>
2. Kennedy, D.A., Lee, T., Seely, D.: A comparative review of thermography as a breast cancer screening technique. *Integr. Cancer Ther.* **8**(1), 9–16 (2009). <https://doi.org/10.1177/1534735408326171>
3. Usamentiaga, R., Venegas, P., Guerediaga, J., Vega, L., Molleda, J., Bulnes, F.: Infrared thermography for temperature measurement and non-destructive testing. *Sensors* **14**(7), 12305–12348 (2014). <https://doi.org/10.3390/s140712305>
4. Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: a brief review. *Comput. Intell. Neurosci.* **2018**, 1–13 (2018). <https://doi.org/10.1155/2018/7068349>
5. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**(1), 221–248 (2017). <https://doi.org/10.1146/annurev-bioeng-071516-044442>
6. Park, G., et al.: Modeling heat transfer in humans for body heat harvesting and personal thermal management. *Appl. Energy* **323**, 119609 (2022). <https://doi.org/10.1016/j.apenergy.2022.119609>
7. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *J Big Data* **6**(1), 60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* (1979), **313**(5786), 504–507 (2006). <https://doi.org/10.1126/science.1127647>
9. Goodfellow, I., et al.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2014). <https://doi.org/10.1145/3422622>
10. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style (2015)
11. Dubey, S.R., Singh, S.K.: Transformer-based generative adversarial networks in computer vision: a comprehensive survey (2023)
12. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. (2017), Accessed: Jan. 18, 2023 [Online]. Available: <https://arxiv.org/abs/1703.10593>
13. Mirza, M., Osindero, S.: Conditional generative adversarial nets (2014)
14. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN (2017)

15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation (2017). <https://doi.org/10.48550/arxiv.1710.10196>
16. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks (2018)
17. Lyu, Y., Xu, Y., Jiang, X., Liu, J., Zhao, X., Zhu, X.: AMS-PAN: breast ultrasound image segmentation model combining attention mechanism and multi-scale features. *Biomed. Signal Process. Control* **81**, 104425 (2023). <https://doi.org/10.1016/j.bspc.2022.104425>
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium (2017). <https://doi.org/10.48550/arxiv.1706.08500>
19. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, June 2019, pp. 658–666. <https://doi.org/10.1109/CVPR.2019.00075>
20. Campos, D.: Métricas de similitud para cadenas de texto. Parte III: Métricas de conjuntos y emparejamiento de caracteres. *SoldAI*
21. Jordan, J.: Evaluating image segmentation models. *Data Sci.* Accessed: Jan 25 2023 [Online]. Available: <https://www.jeremyjordan.me/evaluating-image-segmentation-models/>
22. Betzalel, E., Penso, C., Navon, A., Fetaya, E.: A study on the evaluation of generative models. *Comput. Vis. Pattern Recognit.* (2022)
23. Li, S., Dong, M., Du, G., Mu, X.: Attention dense-U-net for automatic breast mass segmentation in digital mammogram. *IEEE Access* **7**, 59037–59047 (2019). <https://doi.org/10.1109/ACCESS.2019.2914873>
24. Guo, Y., Duan, X., Wang, C., Guo, H.: Segmentation and recognition of breast ultrasound images based on an expanded U-Net. *PLoS ONE* **16**(6), e0253202 (2021). <https://doi.org/10.1371/journal.pone.0253202>
25. Yusoff, M., Abd Karim, M.S.A.: Enhancing histopathological imaging of breast cancer with generative adversarial networks. *J. Comput. Imaging* **30**(3), 205–215 (2020)
26. AlFayez, F., El-Soud, M.W.A., Gaber, T.: Thermogram breast cancer detection: a comparative study of two machine learning techniques. *Appl. Sci.* **10**(2), 551 (2020). <https://doi.org/10.3390/app10020551>
27. Nomani, A., Ansari, Y., Nasirpour, M.H., Masoumian, A., Pour, E.S., Valizadeh, A.: PSOWNNs-CNN: a computational radiology for breast cancer diagnosis improvement based on image processing using machine learning methods. *Comput. Intell. Neurosci.* **2022**, 1–17 (2022). <https://doi.org/10.1155/2022/5667264>



Evaluating Histopathological Cancer Detection: A Comparative Analysis of CNN Architectures for Tumor Detection in Lymph Node Pathology

Ana Marcillo-Vera¹ , Karen Cáceres-Benítez² ,
Diego Almeida-Galárraga²  , and Andrés Tirado-Espín¹ 

¹ School of Mathematical and Computational Sciences, Universidad Yachay Tech, 100119 Urcuqui, Ecuador

² School of Biological Sciences and Engineering, Universidad Yachay Tech, 100119 Urcuqui, Ecuador

dalmeida@yachaytech.edu.ec

Abstract. This study evaluates four well-known convolutional neural networks: VGG19, EfficientNetB0, ResNet50, and InceptionV3, for tumor detection in lymph node pathology images. Using a significant subset of the PCam dataset, models were trained on binary classification tasks focused on identifying tumor tissue, with data augmentation applied to enhance generalization. The methodology involved a rigorous process of data preparation, model selection, training, and evaluation under limited hardware resources, using a standard laptop. The dataset was split into training and validation sets with an 80/20 ratio, and models were trained using the Adam optimizer with a learning rate of 0.001 over multiple epochs. VGG19 achieved the highest validation accuracy at 77.38% and AUC of 85.35% but required substantial computational time and exhibited overfitting. EfficientNetB0, though faster to train (20 m 40 s), showed lower validation accuracy (58.91%) and AUC (60.14%). ResNet50 performed well during training but faced generalization challenges. InceptionV3 demonstrated a balanced performance with a validation accuracy of 70.74% and AUC of 76.41%, making it a promising option across varied datasets. These findings highlight the strengths and limitations of different CNN architectures in enhancing cancer diagnosis in lymph node pathology, providing insights for future research and clinical applications.

Keywords: Metastatic cancer · EfficientNetB0 · VGG19 · ResNet50 · InceptionV3 · Image segmentation · Tumor detection

1 Introduction

The lymphatic system, composed of a vast network of vessels and lymph nodes, is integral to the body's immune function and fluid balance. Lymph nodes act

as critical checkpoints, filtering pathogens and malignant cells, which subsequently activate the immune response [1]. As such, lymphadenopathy, or the enlargement of lymph nodes, can be indicative of various conditions, including infections, inflammatory diseases, and malignancies [2]. The major lymph node groups are strategically located in the body, such as in the neck, axilla, thorax, abdomen, and groin, as shown in Fig. 1. These nodes serve as sentinel sites for detecting pathological changes. In clinical practice, diagnostic techniques such as endoscopic ultrasound (EUS) and elastography have become indispensable tools for the precise detection of malignancies. EUS allows for detailed imaging and fine-needle aspiration of lymph nodes, while elastography provides information on tissue stiffness, which is crucial for identifying malignant transformations. However, despite the technological advancements in these methods, the correlation with clinical findings remains essential to ensure accurate diagnosis and treatment planning [3].

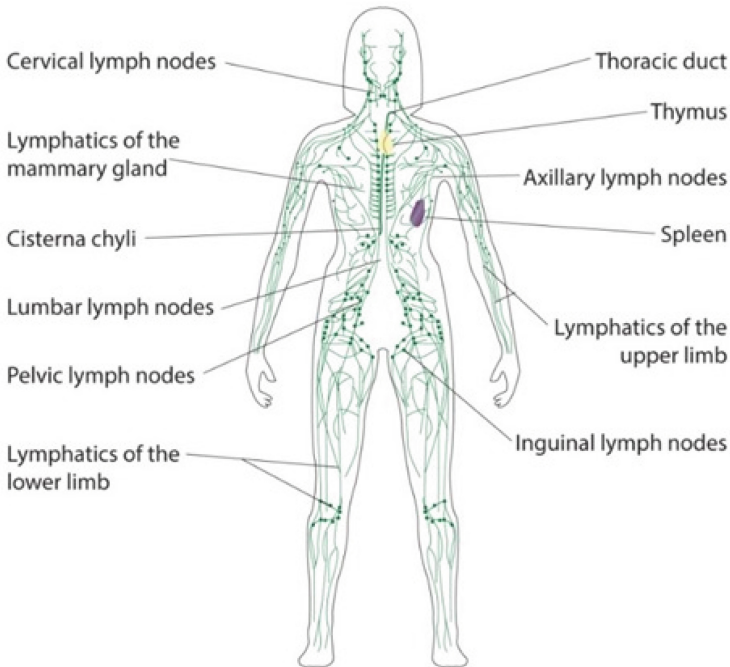


Fig. 1. Major lymph node locations in the human body. Adapted from [4].

Histopathological examination of lymph node tissue remains the gold standard for identifying metastatic cancer [5]. This process involves microscopic analysis of tissue samples to detect cancerous cells. Differentiation between normal and malignant lymph nodes, as illustrated in Fig. 2, is crucial for appropriate

clinical management. Normal lymph nodes typically have a uniform architecture, whereas malignant nodes exhibit disrupted structures and the presence of metastatic cells [3, 6].

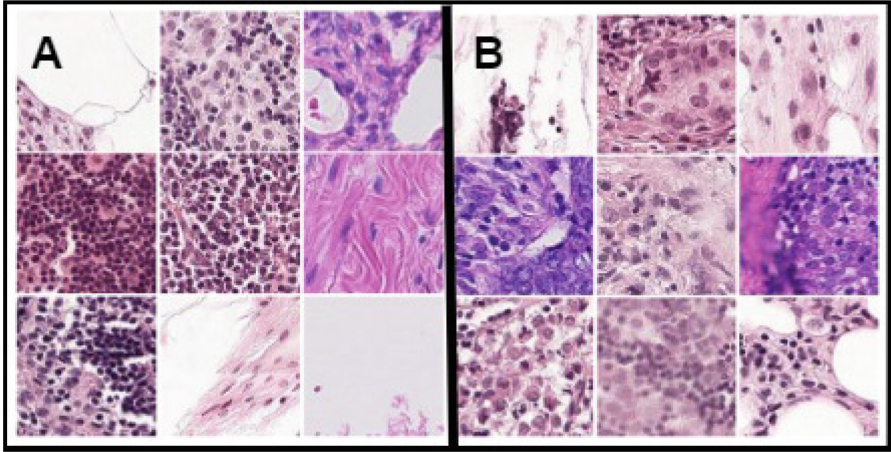


Fig. 2. Comparison of normal (A) and malignant lymph nodes (B). Adapted from [7].

Advances in digital pathology and machine learning offer promising avenues for enhancing cancer diagnosis. In particular, convolutional neural networks (CNNs) have shown significant potential in image analysis tasks [8, 9]. Our study aims to leverage CNNs to develop a binary classification model capable of identifying tumor tissue in digital pathology images of lymph node sections. The choice of 32×32 pixel image size was driven by the availability of a publicly accessible dataset with these specific dimensions, which are well-suited for capturing critical features in small tissue regions. By automating the detection process, we hope to improve the accuracy and speed of cancer diagnosis, thereby facilitating early detection and treatment.

The CNN models evaluated in this study include VGG19, EfficientNetB0, ResNet50, and InceptionV3. Each model has distinct characteristics that influence its performance. VGG19 is known for its depth and simplicity, consisting of 19 layers with small receptive fields of 3×3 convolutional filters [10]. This architecture allows for a high level of precision in feature extraction, making it effective for image classification tasks. However, VGG19 demands significant computational resources, both in terms of memory and processing power, which can be a limitation for large-scale applications. Additionally, due to its depth, VGG19 is prone to overfitting, especially when trained on small datasets without adequate regularization [11]. EfficientNetB0 employs a novel compound scaling method that uniformly scales the network's depth, width, and resolution using a set of fixed scaling coefficients [12]. This approach enables EfficientNetB0 to achieve better performance with fewer parameters compared to traditional

CNNs. One of its main strengths is its efficiency, as it can achieve high accuracy with relatively low computational cost. However, EfficientNetB0 requires meticulous tuning to optimize its performance and mitigate the risk of overfitting, which can be challenging without extensive experimentation and expertise [13].

ResNet50 introduces the concept of residual learning through skip connections, allowing the network to learn residual functions with reference to the input layer [14]. This innovation helps mitigate the vanishing gradient problem and enables the training of much deeper networks. ResNet50, with its 50 layers, excels in training scenarios and has shown strong performance across various benchmarks. Nevertheless, it can struggle with generalization, particularly when applied to datasets that differ significantly from those it was trained on, potentially leading to overfitting if not managed carefully [15, 16]. InceptionV3 uses a sophisticated architecture that incorporates multiple types of convolutional and pooling layers within each module, allowing it to capture various levels of detail [17]. This multi-scale processing capability helps InceptionV3 balance learning capacity and generalization performance, making it a strong candidate for diverse pathology datasets. InceptionV3's ability to process images at different scales and resolutions is one of its key strengths. However, its complex architecture can be computationally expensive and may require extensive hyperparameter tuning to achieve optimal performance [18, 19].

Our study has set forth the following objectives: the main objective is to develop a binary classification model using CNNs to identify tumor tissue in digital images of lymph node sections. Specifically, we aim to compare and evaluate the performance of four CNN architectures (ResNet50, VGG19, InceptionV3, and EfficientNetB0). We also seek to determine the best performing CNN architecture in terms of accuracy, sensitivity, and specificity. To achieve these objectives, we address the following research questions: **RQ1** How do ResNet50, VGG19, InceptionV3, and EfficientNetB0 perform in detecting tumor tissue in lymph node images? **RQ2** What are the accuracy differences among these CNN architectures in tumor tissue detection? **RQ3** Which CNN architecture shows the highest effectiveness in enhancing cancer detection and treatment?

1.1 Related Works

In recent years, significant advancements have been made in the application of deep learning models to predict lymph node metastasis, Table 1. These advancements have shown promising results, both quantitatively and qualitatively, which are relevant to our project.

Lin et al. [20] developed a deep learning model utilizing convolutional neural networks (CNN) to predict axillary lymph node metastasis from ultrasound (US) images. Their model achieved an impressive area under the receiver operating characteristic curve (AUC) of 0.91, demonstrating high diagnostic accuracy. They also employed saliency maps to visualize the regions of interest within the US images, enhancing the interpretability of the model's predictions and providing valuable insights for clinical decision-making.

Table 1. Summary of related works with objectives, models, and evaluation metrics

References	Main Objective	Model	Accuracy	AUC	Sensitivity
[20]	Prediction of axillary lymph node metastasis using CNN on ultrasound images	CNN	-	0.91	-
[21]	Comparison of machine learning algorithms for predicting breast cancer recurrence	Various ML algorithms	82%	0.88	-
[22]	Prediction of axillary lymph node metastasis using deep learning models on ultrasound images	Inception V3, Inception-ResNet V2, ResNet 101	-	0.90 (int.), 0.89 (ext.)	85%
[23]	Using deep learning models to predict axillary lymph node metastasis on ultrasound images	Inception V3	79%	-	85%

Similarly, Tekchandani et al. [21] compared various machine learning algorithms to predict breast cancer recurrence. Their study revealed that deep learning models outperformed traditional methods such as logistic regression and support vector machines, with their deep learning model achieving an accuracy of 82% and an AUC of 0.88. This research highlighted the importance of feature selection and preprocessing steps in enhancing model performance, and underscored the potential of integrating machine learning models into clinical workflows to provide personalized treatment recommendations and improve patient outcomes.

Zhou et al. [22] focused on using deep learning models to predict lymph node metastasis specifically from US images in breast cancer patients. They tested three pretrained CNN models-Inception V3, Inception-ResNet V2, and ResNet 101. The Inception V3 model stood out with an AUC of 0.90 in the internal test set and 0.89 in the external test set. Notably, their study demonstrated that the deep learning models outperformed radiologists in accuracy, sensitivity, and specificity. Moreover, the use of class activation maps provided a visual interpretation of the model's decision-making process, which is crucial for gaining clinicians' trust and facilitating the adoption of these models in clinical practice.

Bae [23] also conducted a study to predict axillary lymph node metastasis using deep learning models trained on US images of breast cancer patients. Their models, including Inception V3, Inception-ResNet V2, and ResNet 101, showed superior performance compared to radiologists. The Inception V3 model, in particular, achieved an accuracy of 79%, a sensitivity of 85%, and a specificity of 73% in the external test set. Bae et al. emphasized the potential of deep learning models to reduce interobserver variability and improve diagnostic consistency. They also discussed the importance of incorporating clinical-pathologic factors into the models to enhance predictive accuracy and support personalized medicine.

These studies collectively demonstrate the substantial potential of deep learning models in predicting axillary lymph node metastasis and breast cancer recurrence. The quantitative metrics, including high AUC, accuracy, sensitivity, and specificity, underscore the superior performance of deep learning models over traditional methods and radiologists. Qualitatively, the use of visualization techniques such as saliency maps and class activation maps enhances model interpretability, which is crucial for clinical adoption. Integrating these advanced models into clinical workflows can significantly improve diagnostic accuracy, reduce variability, and support personalized treatment planning, ultimately leading to better patient outcomes.

In comparison to other architectures not included in our study, DenseNet has demonstrated its robustness in medical image classification. Its dense connection strategy enhances feature reuse across the network, leading to improved accuracy. Recent studies show that DenseNet outperforms classical models like ResNet34 and VGG19, particularly in identifying metastatic cancer in small image patches from larger digital pathological scans [24]. On the other hand, MobileNetV3, known for its efficiency and compactness, achieved high accuracy and recall in classifying breast histopathology images, though its performance varied across different histological subcategories [25]. These findings suggest that while our chosen models balance performance and computational efficiency, future work should explore DenseNet and MobileNetV3 for potential advancements in lymph node pathology image classification.

2 Methodology

This section outlines the methodology employed in this study to classify pathology images using various convolutional neural network (CNN) architectures. Figure 3 illustrates the overall process, which includes data preparation, model selection, training, and evaluation. The subsequent sections provide detailed descriptions of each step, ensuring that the approach is both rigorous and reliable. The goal of this study is to leverage advanced deep learning techniques to improve the accuracy and robustness of pathology image classification, which is crucial for early and accurate diagnosis of diseases. The following section provide a detailed description of each methodological step, along with justifications and references to pertinent scientific research.

2.1 Data Description and Computational Environment

This study utilizes a dataset of small pathology images, each identified by a unique image ID. The ground truth labels for the central 32×32 px region of each image are provided in the `train_labels.csv` file, indicating whether tumor tissue is present (1) or absent (0). This dataset is a modified version of the PCam dataset, designed to remove duplicate images while retaining the original data splits, as per the PCam benchmark [26]. The removal of duplicates ensures an

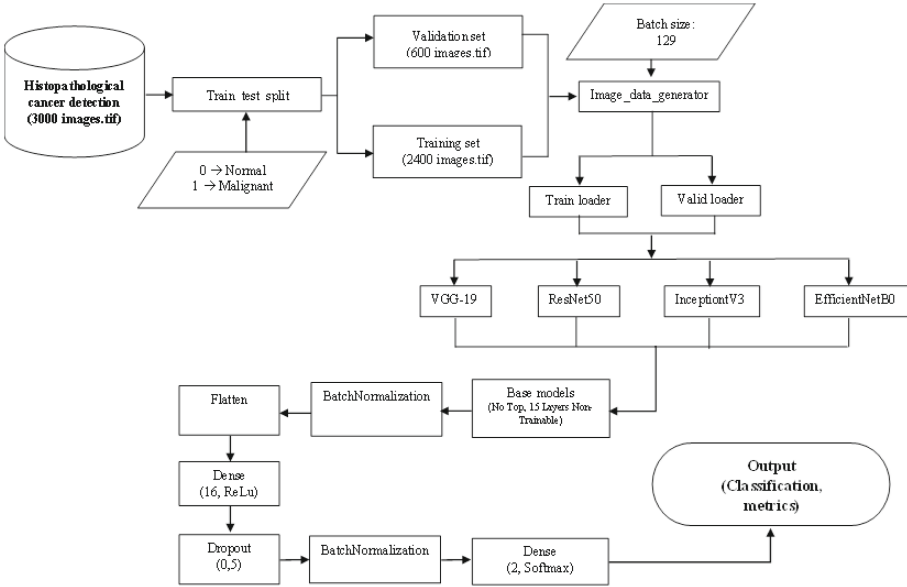


Fig. 3. Flowchart of methodology employed

unbiased model training process, leading to a more accurate evaluation of model performance.

The analysis was conducted on a Dell Inspiron laptop with a 7th generation Intel Core i5 processor, using the Jupyter environment for coding and data analysis. Due to computational limitations, a subset of 3000 images was selected, sufficient for effective model training and evaluation. This setup highlights the feasibility of employing deep learning techniques on personal laptops, making advanced research accessible to those without high-performance computing resources.

2.2 Data Splitting

To ensure a robust evaluation of the models, the dataset was split into training and validation sets with an 80/20 split ratio, where 80% of the data (2400 images) was allocated for training and 20% (600 images) was reserved for validation. This split ratio is a widely accepted practice in machine learning to balance having enough data to train the model and enough data to validate its performance [27]. However, a more comprehensive study on partition sizes, including alternative split ratios (e.g., 70/30 or 90/10), could provide additional insights into how different partition sizes impact model training and evaluation. Exploring these variations can help determine if the chosen split ratio is optimal or if adjustments might lead to improved performance or generalization [28]. This approach ensures that the model's performance metrics are a true reflection of

its ability to classify unseen data accurately, while also evaluating the robustness of the model across different data partitions.

2.3 Batch Size Selection

A batch size of 129 was chosen for this study, balancing the need for efficient computation and stable training. Objectively, a batch size of 129 was relatively large for the system used, but it was large enough to benefit from stable gradient updates but still small enough to fit into the available memory without causing excessive computational load. With 2400 training images, this batch size allows the model to process almost 20% of the training data in each batch, leading to fewer gradient updates per epoch. This can result in more stable training dynamics and quicker convergence, which is particularly beneficial when dealing with a smaller dataset where each update has a significant impact on the model's learning. However, batch sizes that are powers of two are often preferred due to their alignment with the memory access patterns of GPUs, which can lead to optimized performance [29]. The selected batch size maximized the balance between training speed and stability, ensuring effective use of available computational resources.

2.4 Data Augmentation

To enhance the model's robustness and prevent overfitting, data augmentation was performed using the ImageDataGenerator function from Keras. Data augmentation techniques, such as rotation, shifting, zooming, and flipping, were applied to artificially expand the training dataset. These techniques help the model learn invariance to various transformations, thereby improving its generalization capabilities [30]. By exposing the model to a wider variety of altered images during training, we aim to create a more robust model that performs well on real-world data. Data augmentation is particularly useful in medical imaging, where acquiring large amounts of labeled data can be challenging. It allows us to make the most of the available data by simulating additional training examples through realistic transformations [30].

2.5 Model Selection, Architecture, and Training

This study utilized four CNN architectures—VGG19, ResNet50, InceptionV3, and EfficientNetB0—chosen for their strong performance in image classification and availability in pre-trained forms [31–34]. Pre-trained weights from ImageNet were used, with the first 15 layers of each model frozen to retain low-level features while fine-tuning the remaining layers for tumor classification [35]. Training epochs were adjusted to suit each model's characteristics, as detailed in Table 2.

The architecture of each model included batch normalization for stable training, a flatten layer to convert 2D data into a 1D vector, a dense layer with 16 units and ReLU activation for non-linearity, and a dropout layer (rate of 0.5) to

Table 2. Model Architectures Used

Model	Pre-trained Weights	Frozen Layers	Epochs Used
VGG19	ImageNet	15	10
ResNet50	ImageNet	15	10
InceptionV3	ImageNet	15	20
EfficientNetB0	ImageNet	15	30

prevent overfitting [36]. The final layer used softmax activation to classify the images as normal or tumor.

Training was performed using the Adam optimizer with a 0.001 learning rate, known for its efficiency in complex parameter spaces [37]. The categorical cross-entropy loss function was applied, and models were trained for 50 epochs with early stopping based on validation loss to prevent overfitting. Performance was assessed using accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of the models' effectiveness.

3 Results

This section presents the results of the classification task, detailing the performance of each convolutional neural network (CNN) architecture employed. The evaluation metrics include accuracy, area under the curve (AUC), and loss, providing a comprehensive assessment of the models' effectiveness in distinguishing between normal and tumor tissues. The models evaluated in this study were VGG19, ResNet50, InceptionV3, and EfficientNetB0. Each model was trained and validated using the methodology described, with the results summarized in Table 3 and illustrated in Figs. 4, 5, and 6.

3.1 Accuracy and AUC

Figures 6 and 5 illustrate the training and validation accuracies and AUCs, respectively, of the different CNN architectures. ResNet50 and EfficientNetB0 demonstrated the highest training accuracies at 0.8514 and 0.8429, respectively. However, EfficientNetB0's validation accuracy was lower than expected at 0.5891, suggesting potential overfitting. VGG19 showed balanced performance with training and validation accuracies of 0.7674 and 0.7738, respectively. The

Table 3. Performance metrics of CNN architectures in lymph node tumor detection

Model	Training Accuracy	Training AUC	Training Loss	Validation Accuracy	Validation AUC	Validation Loss	CPU Time
VGG19	0.7674	0.8658	0.4814	0.7738	0.8535	0.5343	54 m 57 s
EfficientNetB0	0.8429	0.9293	0.3569	0.5891	0.6014	0.6888	20 m 40 s
ResNet50	0.8514	0.9320	0.3594	0.6008	0.5896	0.6783	30 m 32 s
InceptionV3	0.6518	0.7281	0.6548	0.7074	0.7641	0.6344	35 m 17 s

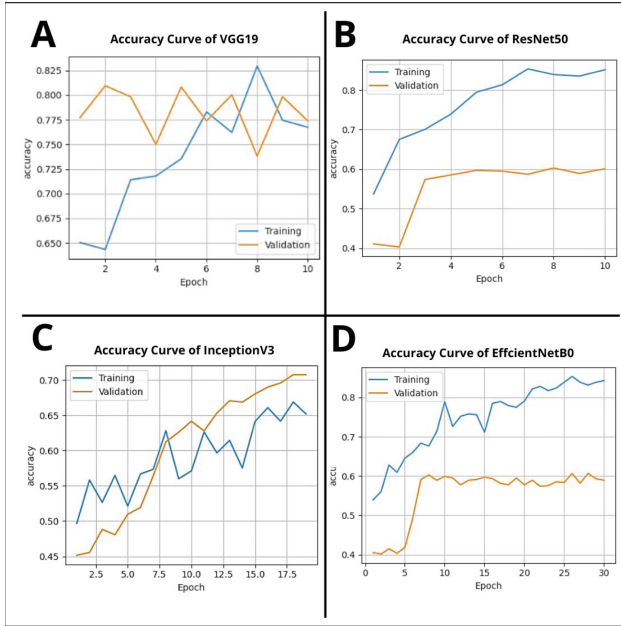


Fig. 4. Training and Validation Accuracy Curves of CNN Architectures: (A) VGG19, (B) ResNet50, (C) InceptionV3, (D) EfficientNetB0

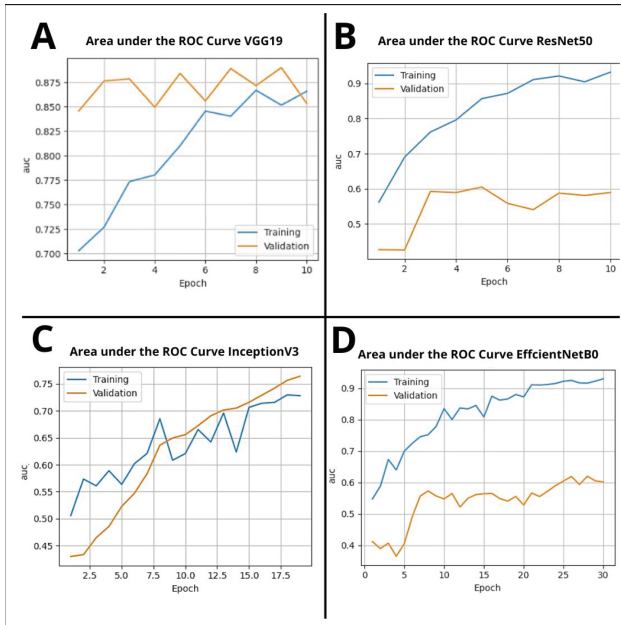


Fig. 5. Training and Validation AUC Curves of CNN Architectures: (A) VGG19, (B) ResNet50, (C) InceptionV3, (D) EfficientNetB0

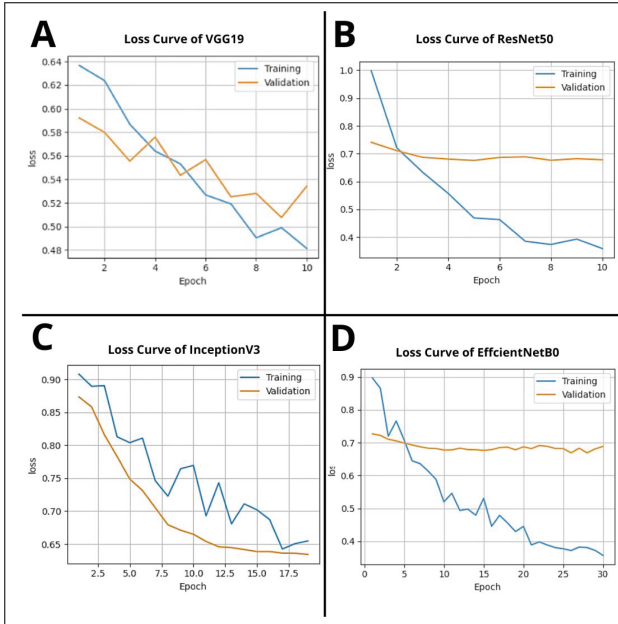


Fig. 6. Training and Validation Loss Curves of CNN Architectures: (A) VGG19, (B) ResNet50, (C) InceptionV3, (D) EfficientNetB0

AUC values followed a similar trend, with ResNet50 achieving the highest training AUC at 0.9320 and VGG19 showing a strong validation AUC of 0.8535, indicating good model generalization.

3.2 Loss Metrics and Computational Efficiency

Figure 4 provides an overview of the training and validation losses. ResNet50 and EfficientNetB0 exhibited the lowest training losses, indicating efficient learning. However, EfficientNetB0's higher validation loss (0.6888) compared to its training loss (0.3569) highlights the need for better regularization to improve generalization. In contrast, VGG19 showed more consistent loss metrics, with a training loss of 0.4814 and a validation loss of 0.5343. InceptionV3's training loss was the highest at 0.6548, but its validation loss of 0.6344 indicates that it performed reasonably well despite its lower accuracy metrics. In terms of computational efficiency, EfficientNetB0 outperformed the other models with a CPU time of 20 min and 40 s, compared to VGG19's 54 min and 57 s, demonstrating its efficiency despite its complexity. On the other hand, VGG19 and ResNet50, both trained for 10 epochs, provided robust performance metrics, suggesting that a moderate number of epochs is sufficient for these architectures. InceptionV3, trained for 20 epochs, showed improvements in validation metrics, indicating the benefits of extended training for more complex models. EfficientNetB0, with

30 epochs, achieved high performance in CPU time but also exhibited signs of overfitting, as seen in its lower validation accuracy and higher validation loss.

4 Discussion

The evaluation of various convolutional neural network (CNN) architectures VGG19, ResNet50, InceptionV3, and EfficientNetB0-yielded insightful results regarding their performance in pathology image analysis. Each model was assessed using metrics such as accuracy, area under the curve (AUC), and computational efficiency, providing a comprehensive view of their strengths and considerations for practical deployment.

ResNet50 consistently demonstrated strong performance metrics, achieving the highest training accuracy (0.8514) and training AUC (0.9320) among the models. This indicates its capability to effectively learn and generalize complex patterns in the training data, which is critical for accurate prediction in pathology image classification tasks. VGG19 followed closely with competitive validation metrics (accuracy: 0.7738, AUC: 0.8535), highlighting its robustness and reliability in capturing relevant features from the data. EfficientNetB0, known for its computational efficiency, exhibited the lowest training loss, emphasizing its effective utilization of training data. However, its validation accuracy (0.5891) and AUC (0.6014) were notably lower, suggesting potential overfitting issues that may require regularization techniques for improved generalization. InceptionV3, despite showing higher training loss, demonstrated balanced validation metrics (accuracy: 0.7074, AUC: 0.7641), indicating its adaptability and resilience in capturing meaningful features from pathology images. EfficientNetB0 emerged as the most efficient model in terms of CPU time (20 min and 40s), making it suitable for deployment in scenarios requiring rapid processing or resource-constrained environments.

The results underscore the strengths of ResNet50 and VGG19 in pathology image classification, offering high accuracy and reliable generalization across different datasets. These models are well-suited for applications demanding precise diagnostic capabilities and disease prognosis. EfficientNetB0, while efficient in training, requires strategies to address overfitting observed in validation metrics. Techniques such as regularization methods (e.g., dropout) or increased data diversity could enhance its performance and reliability in real-world applications. InceptionV3 presents a balanced option with moderate computational demands and reasonable accuracy, making it a practical choice for scenarios prioritizing a balance between model performance and resource efficiency. When comparing our study with previous works, our research primarily emphasizes the assessment of accuracy and AUC metrics. These metrics are crucial for evaluating the CNN models' ability to effectively distinguish between normal and tumor tissues. Notably, while other studies such as [21–23] reported sensitivity metrics alongside accuracy and AUC, our study did not include sensitivity metrics in the evaluation. This distinction highlights the emphasis placed on different performance indicators across various studies in the field of medical image analysis. Regarding accuracy, our study achieved competitive results with VGG19

and ResNet50, ranging from 76.74% to 85.14%. These accuracies are comparable to those reported by [21], who attained 82% using various machine learning algorithms for breast cancer recurrence prediction. This suggests that our CNN models effectively discriminate between normal and tumor tissues, performing similarly or better than traditional machine learning methods in similar tasks. In terms of AUC, our study showed strong performance, particularly with ResNet50, which achieved a training AUC of 93.20% and a validation AUC of 58.96%. This aligns with the findings of [22], who reported AUC values of 90% (internal) and 89% (external) using Inception V3, Inception-ResNet V2, and ResNet 101 models. This consistency in AUC metrics underscores the robustness of deep learning approaches in handling the complexities of ultrasound image classification for lymph node metastasis detection. EfficientNetB0, despite showing slightly lower validation accuracy and higher validation loss compared to other models in our study, demonstrated efficient learning with a training AUC of 92.93% and a validation AUC of 60.14%. Its computational efficiency, with a reduced training time of 20 min and 40 s, makes it a viable option for applications requiring rapid diagnostic processing.

In addressing **RQ1**, the performance of ResNet50, VGG19, InceptionV3, and EfficientNetB0 in detecting tumor tissue in lymph node images varied significantly across different metrics. ResNet50 and VGG19 showed strong training performance, with accuracies of 0.8514 and 0.7674, respectively, while InceptionV3, despite being trained for the most epochs, maintained a more balanced performance across training and validation, indicating better generalization. EfficientNetB0, although it displayed the shortest training time (20 min and 40 s), exhibited significant overfitting, with a substantial drop in validation accuracy (0.5891) compared to its training accuracy (0.8429). Regarding **RQ2**, the accuracy differences among these CNN architectures were notable, with ResNet50 achieving the highest training accuracy but suffering from overfitting during validation (0.6008), while VGG19 exhibited the highest validation accuracy (0.7738). However, VGG19 required significantly more training time (54 min and 57 s), which may not be favorable for practical applications where computational efficiency is critical. InceptionV3, though having lower overall accuracy, demonstrated more consistent results between training and validation, highlighting its robustness. For **RQ3**, although ResNet50 and VGG19 showed strong detection capabilities, InceptionV3 emerged as the most effective architecture for enhancing cancer detection and treatment due to its balanced performance, lower degree of overfitting, and reasonable training time (35 min and 17 s), making it a potentially more reliable model in clinical applications.

When considering the broader implications and applications, our study's results underscore the effectiveness of CNN architectures in pathology image analysis. The ability to achieve high accuracy and AUC metrics demonstrates the potential of deep learning models to enhance diagnostic accuracy and efficiency in clinical settings. These findings contribute to advancing the use of AI in healthcare by providing reliable tools for early disease detection and treatment planning. In comparison with related works, our study contributes by focusing on

a comprehensive evaluation of multiple CNN architectures specifically for lymph node tumor detection from ultrasound images. While some studies may report additional metrics like sensitivity, our emphasis on accuracy and AUC aligns with current standards in medical image analysis and highlights the strengths of CNN models in this domain.

5 Conclusion

Our study aimed to develop a binary classification model using convolutional neural networks (CNNs) to identify tumor tissue in digital images of lymph node sections. We evaluated the performance of four CNN architectures: ResNet50, VGG19, InceptionV3, and EfficientNetB0, to determine the best-performing model in terms of accuracy, sensitivity, and specificity. ResNet50 and VGG19 demonstrated the highest effectiveness in detecting tumor tissue. ResNet50 achieved a training accuracy of 0.8514 and a validation accuracy of 0.6008, while VGG19 had a training accuracy of 0.7674 and a validation accuracy of 0.7738. EfficientNetB0, despite high training accuracy (0.8429) and AUC (0.9293), showed lower validation accuracy (0.5891) and higher validation loss (0.6888), indicating potential overfitting. InceptionV3, trained for the most epochs, displayed balanced but lower performance metrics compared to ResNet50 and VGG19. However, visualizations of training and validation curves highlighted InceptionV3's superior ability to maintain a stable balance between training and validation metrics, with less divergence between them, which indicates a lower degree of overfitting. Although the accuracy achieved was not as high as expected for a trustworthy application, it is crucial to consider how well-balanced a model is. For instance, while EfficientNetB0 and ResNet50 exhibited high accuracies above 80%, they encountered severe overfitting issues during validation. This highlights the importance of achieving a balance between accuracy and model generalization to ensure reliable performance in practical applications.

On the other hand, the experimentation faced notable limitations due to the small dataset used to avoid exceeding computational constraints, which may have affected the generalizability of the findings. According to the dataset specifications, all images utilized in the study were acquired with the same sensor. Although our analysis showed that data augmentation techniques improved model robustness, further exploration is necessary to evaluate the model's performance with images captured from different sensors or camera sources. Assessing how the model performs on data from varied acquisition conditions could provide valuable insights into its generalization capabilities and overall robustness across diverse imaging environments.

Future work should focus on evaluating the model's performance with new tumor images from various sensors or camera sources. Although validation with unseen images was conducted, assessing the model's generalization to data from different imaging conditions could provide deeper insights into its robustness. Testing and training with images from diverse sources might reveal whether such data influences model performance more than augmentation alone. Additionally, expanding the dataset, fine-tuning CNN architectures, and implementing

cross-validation techniques will help improve model generalization and mitigate overfitting. Investigating advanced preprocessing methods and developing practical diagnostic applications will further enhance the model's clinical utility and effectiveness in cancer detection.

References


1. Hantusch, B.: Morphological and functional characteristics of blood and lymphatic vessels. In: *Fundamentals of Vascular Biology* (2019), pp. 1–43
2. Ahmed, Z., et al.: Kikuchi-Fujimoto disease: a rare cause of cervical lymphadenopathy. *Cureus* **13**(8) (2021)
3. Tamanini, G., Cominardi, A., Brighi, N., Fusaroli, P., Lisotti, A.: Endoscopic ultrasound assessment and tissue acquisition of mediastinal and abdominal lymph nodes. *World J. Gastrointest. Oncol.* **13**(10), 1475 (2021)
4. Staff, B.: Medical gallery of Blausen medical 2014. *Wiki J. Med.* **1**(2) (2014)
5. Tseng, L.J., Matsuyama, A., MacDonald-Dickinson, V.: Histology: the gold standard for diagnosis? *Can. Vet. J.* **64**(4), 389 (2023)
6. Hu, Y., Su, F., Dong, K., Wang, X., Zhao, X., Jiang, Y., Li, J., Ji, J., Sun, Y.: Deep learning system for lymph node quantification and metastatic cancer identification from whole-slide pathology images. *Gastric Cancer* **24**, 868–877 (2021)
7. Cukierski, W.: *Histopathologic cancer detection* (2018)
8. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., Abdel-Mottaleb, M.: Convolutional neural networks for breast cancer detection in mammography: a survey. *Comput. Biol. Med.* **131**, 104248 (2021)
9. Sibille, L., et al.: 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology* **294**(2), 445–452 (2020)
10. Sushma, L., Lakshmi, K.: An analysis of convolution neural network for image classification using different models. *Int. J. Eng. Res. Technol. (IJERT)* **9**(10) (2020)
11. Bansal, M., Kumar, M., Sachdeva, M., Mittal, A.: Transfer learning for image classification using VGG19: Caltech-101 image data set. *J. Ambient Intell. Hum. Comput.* (2023) 1–12
12. Patel, C.H., Undaviya, D., Dave, H., Degadwala, S., Vyas, D.: EfficientNetB0 for brain stroke classification on computed tomography scan. In: *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, pp 713–718 (2023)
13. Sruthi, S., Emadaboina, S., Machavarapu, P., Singh, R.P., Kanchan, S.: Covid-19 classification using fine-tuned EfficientNet architecture. In: *IEEE 9th International Conference for Convergence in Technology (I2CT)*. IEEE 2024, pp. 1–6 (2024)
14. Koonce, B., Koonce, B.: Resnet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pp. 63–72 (2021)
15. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. *IEEE Trans. Med. Imaging* **38**(9), 2092–2103 (2019)
16. Gour, M., Jain, S., Sunil Kumar, T.: Residual learning based CNN for breast cancer histopathological image classification. *Int. J. Imaging Syst. Technol.* **30**(3), 621–635 (2020)

17. Wang, C., et al.: Pulmonary image classification based on Inception-V3 transfer learning model. *IEEE Access* **7**, 146533–146541 (2019)
18. Guan, Q., et al.: Deep convolutional neural network Inception-V3 model for differential diagnosing of lymph node in cytological images: a pilot study. *Ann. Transl. Med.* **7**(14) (2019)
19. Sam, S.M., et al.: Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogleNet inception-v1 and inception-v3. *Proc. Comput. Sci.* **161**, 475–483 (2019)
20. Lin, Y.H., et al.: Development and validation of a 3D ResNet model for prediction of lymph node metastasis in head and neck cancer patients. *J. Imag. Inform. Med.* **37**(2), 679–687 (2024)
21. Tekchandani, H., Verma, S., Londhe, N.: Performance improvement of mediastinal lymph node severity detection using GAN and Inception network. *Comput. Methods Programs Biomed.* **194**, 105478 (2020)
22. Zhou, L.Q., Wu, X.L., Huang, S.Y., Wu, G.G., Ye, H.R., Wei, Q., et al.: Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology* **294**(1), 19–28 (2020)
23. Bae, M.S.: Using deep learning to predict axillary lymph node metastasis from US images of breast cancer (2020)
24. Zhong, Z., Zheng, M., Mai, H., Zhao, J., Liu, X.: Cancer image classification based on densenet model. *J. Phys: Conf. Ser.* **1651**, 012143 (2020)
25. DeVoe, K., Takahashi, G., Tarshizi, E., Sacker, A.: Evaluation of the precision and accuracy in the classification of breast histopathology images using the MobileNetV3 model. *J. Pathol. Inform.* **15**, 100377 (2024)
26. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, Springer, pp 210–218 (2018)
27. Chollet, F.: *Deep learning with Python*. Simon and Schuster (2021)
28. Mahmud, M.S., Huang, J.Z., Salloum, S., Emara, T.Z., Sadatdiyov, K.: A survey of data partitioning and sampling methods to support big data analysis. *Big Data Min. Anal.* **3**(2), 85–101 (2020)
29. Goodfellow, I., Bengio, Y., Courville, A.: *Deep learning*. MIT Press, New York, NY (2016)
30. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *J. Big Data* **6**(1), 1–48 (2019)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
32. 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 (2016)
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826 (2016)
34. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, PMLR, pp. 6105–6114 (2019)
35. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? *Adv Neural Inform Process Syst* **27** (2014)

36. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
37. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)



ViTSigat: Early Black Sigatoka Detection in Banana Plants Using Vision Transformer

Jorge L. Charco^(✉) , Angela Yanza-Montalvan, Johanna Zumba-Gamboia, Jose Alonso-Anguizaca, and Edgar Basurto-Cruz

Universidad de Guayaquil, Delta and Kennedy Av., Guayaquil, Ecuador
{jorge.charcoa, angela.yanzam, johanna.zumbag, jose.alonsoan, edgar.basurtoc}@ug.edu.ec

Abstract. This paper presents a model based on Vision Transformer for black Sigatoka detection in early stages, which is called ViTSigat. A comparative of the performance with other proposed model based on Convolutional Neural Network is performed. As first step, a pre-processing is applied to a dataset of 98 videos, which were recorded from a banana plantation using a mobile phone with high resolution camera. The obtained images from each video generated a total of 1500 images, which were normalize and divided in four stages/categories. Different setting parameters were used to train ViTSigat and CNN models. Likewise, both models used the same optimizer, loss function and learning rate. The obtained results were evaluated considering the accuracy, precision, recall and f1-score metrics, including the confusion matrix and ROC curves. A saliency map is used to show the relevant areas where the leaves could be affected due to the infection by black Sigatoka. The experimental results of the proposed models show that the ViTSigat model based on transformer encoder obtains better performance since its attention modules focus on important features of input data instead of take attention to non-useful information.

Keywords: Vision Transformer · Convolutional Neural Network · Black Sigatoka · Attention modules

1 Introduction

Banana farming is one of most important activities in some Latin American countries, having an important role in socio-economic in developing countries. This activity can be performed in all tropical regions. According to mentioned by Food and Agriculture Organization of the United Nations (FAO) [9], the banana production in Latin America and the Caribbean reached 14.5 M tons in 2022 and could reach 36 M tons in 2030. Ecuador is the world's leading producer and exporter of bananas with a production of 5.9M tons. In [7], Continuous Agricultural Surface and Production Survey (ESPAC) published that the banana production in Ecuador increased to 7.2 M tons in 2023.

Currently, agricultural producers are searching strategies that can help to check large hectares of banana plantations to avoid losses of production due to different banana plant diseases, between them, black Sigatoka. A strategy is the inclusion of technological devices such as drones, which can be used to fumigate plantations. However, it is important to first recognize where the banana plantation has been affected to focus the spraying, and thus reduce costs.

Black Sigatoka disease can reduce the quality of the product, or even, increasing the probability of losing 100% of banana production, mainly when the control performed by the agronomist are deficient. This disease is also known as black leaf streak, which has a bunch of intimately correlated fungi. When this infection affects a large part of the leaves, the plants die due to interrupts performing photosynthesis. In [1], the authors have indicated that this infection destroys leaf tissue, doing that visual patterns are presented. This can be leveraged by computer vision techniques to develop new methods to detect this type of disease. The authors in [23] have proposed to use Gaussian filters to soften images, including the usage of thresholds to identify patterns of banana plant diseases considering different visual representation systems such as HSV, TSL, LAB, and YCbCr. However, these proposals are validated under a controlled environment where lighting conditions are constants, doing that thresholds images are not optimal.

In past years, other proposals had used classic machine learning techniques to extract important features of images, and thus, classify banana plantation diseases. In [26], the authors have proposed to use the space color YCbCr considering threshold operations together with supervised learning models such as support vector machines, to classify images of leaves infected with black Sigatoka. Likewise, other techniques as wavelet transform, Fourier transform, SIFT, and SURF [4, 17, 24, 27] have been used to extract important features, and thus, train different architectures such as k-neighborhoods and neural networks [6]. Other proposed approach in [18], the authors have used a k-Fold cross-validation method for detecting banana plant diseases such as banana bacterial wilt (BBW) and banana black Sigatoka (BBS). In this proposal different classifiers were considered such as Nearest Neighbors [16], Decision tree [13], Random forest [5], Extremely Randomized Trees [10] and support vector classifier [11, 12]. The obtained results by Extremely Randomized are better than other classifiers to identify disease BBS. Other work was proposed in [2], where the authors had used an artificial neural network to detect leaf disease and classification of the disease. For this, the color and HOT (Histogram of Template) of images are extracted to train the proposed approach, and then, the total percentage of the affected area is considered to classify the image by its disease type. However, these proposed approaches are only limited to detect the presence of these diseases when a large part of the leaves of banana plant have been already infected, hence, they are not useful to detect banana plant diseases (i.e., black Sigatoka) in early stage, becoming an important challenging to be solved since they could help to perform effective control of this disease. Considering the detection in early stage of black Sigatoka, the authors in [14] have proposed to use a SVM classifier with radial basis function (RBF) kernels, where the images used are close-range hyper-spectral remote sensing images. In order to extract spectral-spatial fea-

tures from banana leaves at both earlier and late stages, Morphological openings and closings are used, which helped to improve time-series hyper-spectral images analysis, getting better predictions of black Sigatoka in early stage (i.e., early detection, mid detection and late detection).

During the last years, the usage of Convolutional Neural Networks (CNNs) have become a powerful tool for image processing, being used in precision agriculture applications such as leaf and stem counts, plant recognition just to mention a few tasks. In this address, the authors in [1] have been proposed to use LeNet architecture to classify Banana Leaf Diseases (i.e., healthy, black Sigatoka and black speckle) using RGB images and the grayscale images, showing appealing results even under challenging conditions such as illumination, complex background, different resolution, size, pose, and orientation. In [20], the authors have proposed to use a pre-trained ResNet50 model to classify healthy or unhealthy banana leaf with black Sigatoka disease. The used images are of 112×112 pixel (RGB channels). Data augmentation was considered to increase the number of images during the training phase. This approach achieves 96.4% of overall accuracy. A similar work is presented in [22], where the authors have utilized the ResNext50 model to classify the Seriousness levels of Sigatoka illness in banana leaves. The dataset contains 10k images, which are labeled using five seriousness levels for training process of proposed model. The model obtained an overall accuracy of 95.53%. In [19], the authors have presented a CNN model to identify leaves with Sigatoka leaf spots and healthy leaf of banana plant. The proposal considered 2k images where 1.6k were used for training process and the remainder was used for testing. The dataset contained enormous, erratic swaths of desiccated tissue, including rusty brown to black patches encircled by a yellow halo, which are the first signs of the condition. The obtained result showed that disease detection accuracy of the intended model was 96.41%.

In the current work, to our knowledge, it is first approach that use Vision Transformer (ViT) architecture for black Sigatoka detection in early stage. The proposed architecture divides the images in small patches to pay more attention to important features of image into learning process, and thus, improve the results. This research has the following contributions:

- Generate a dataset of leaves of banana plants considering early stages of infections process of black Sigatoka (four stages), which are used for training process of proposed model.
- Develop two architectures for black Sigatoka detection in early stages. First, a ViT based architecture and second, a CNN based architecture.
- Compare the experimental results of proposed model with obtained results of CNN to show that the approach based on ViT models offer better performance.

The remainder of the paper is organized as follows. In Sect. 2, the material and methods used in this article are presented, including dataset, proposed models and metrics used in this study. Experimental results are reported in Sect. 3, the discussion of obtained results is presented in Sect. 4. Finally, conclusions are given in Sect. 5



Fig. 1. Farm of banana plantation located in Marcabeli, province *El Oro*.

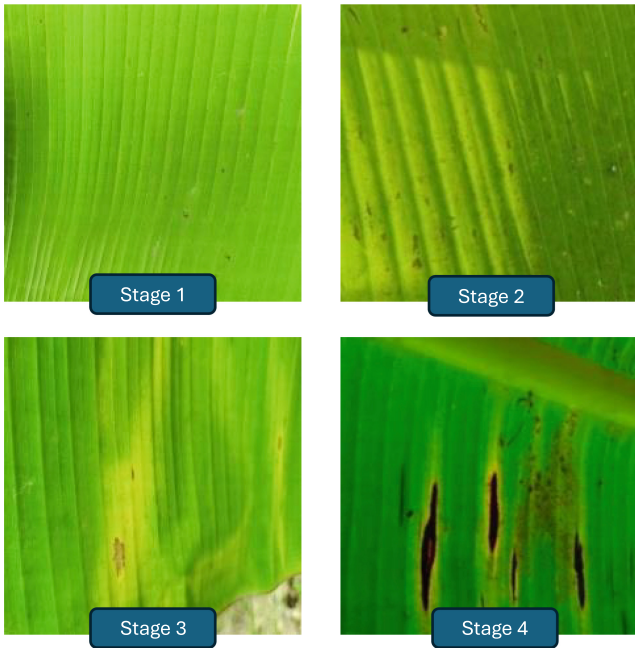


Fig. 2. Leaves of banana plant: Early stage of black Sigatoka.

2 Material and Methods

2.1 Dataset Collection

The dataset for the proposed approach consists of around 1500 images of leaves of banana plants, both healthy and infected by black Sigatoka. For this, 98 videos of banana plants were collected from a banana plantation located in Marcabeli province *El Oro* (see Fig. 1). Each video had a duration of around one minute, which was captured using a mobile phone with high resolution camera sensor (1440p@30fps) considering a distance of 30 cm respect to leaves of banana plant.

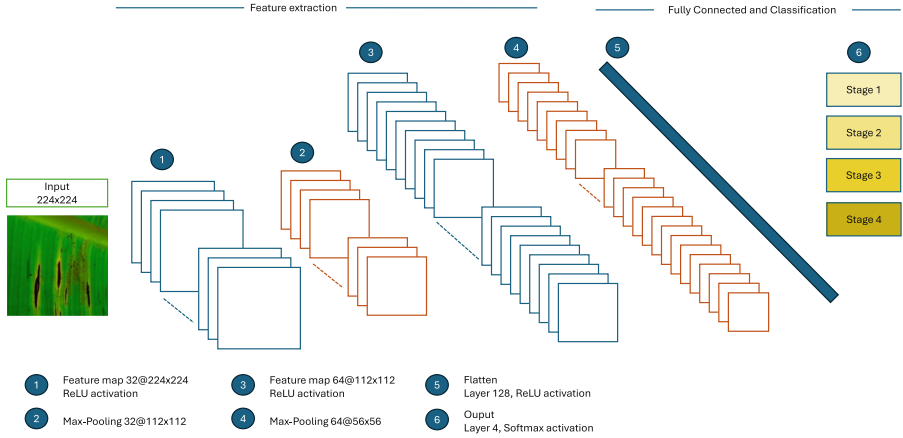


Fig. 3. CNN architecture for black Sigatoka detection

2.2 Image Pre-processing

As pre-processing dataset, the images were extracted from each video recorded previously, generating a total of 1500 images, which were random cropped and resized to 224×224 pixels for training process. Likewise, to normalize the images, the mean value of intensity of pixels and standard deviation were computed and subtracted in each image. The pre-processing mentioned above has also been used during the evaluation phase, with the difference that a center crop on images is used instead of a random crop. According to the obtained images, they were classified in four early stages of infections process of black Sigatoka (see Fig. 2).

2.3 Methods

To perform the black Sigatoka detection, two architectures were proposed with the goal of detecting this disease in early stages. The images dataset used by these architectures for solving this task contains three channels (RGB). The architectures are described below.

Convolutional Neural Network. It is a class of feed-forward neural network, which can learn feature of images using filters (i.e., kernel) optimization. The CNNs are used task with images, speech or audio signal inputs, and they have three types of layers: Convolutional layer, pooling, fully connected layer; being the first, the core building block of a CNN since it is where the majority of computation occurs. Usually, the input corresponds to RGB image, hence, three dimensions are required (width, height and depth). Likewise, a feature detector is required, which is known as kernel or filter, and move across the receptive fields of the image, checking if the feature is present.

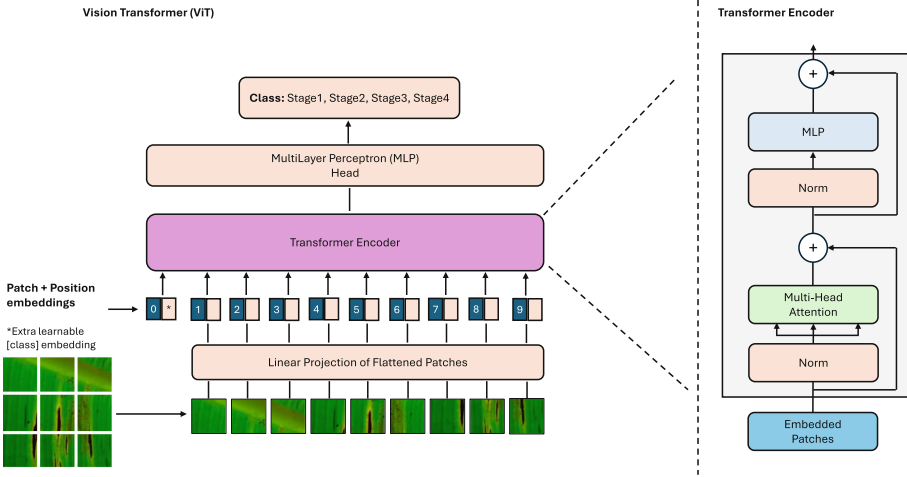


Fig. 4. Transformer architecture: The image is divided into fixed-size patches, which are flattened; and together to the position embeddings are feed to a standard transformer encoder. The illustration of the Transformer encoder was inspired by [25].

The proposed architecture consists of two convolutional layers, two max-pooling and fully-connected (see Fig. 3). Both convolutional layers used a kernel size of 3×3 . Additional, ReLU activation function was considered to introduce a non-linearity to the network. A max-pooling process is also performed, which consists to calculate the maximum activation value for each sub-windows to create a feature map, reducing the spatial dimensions of an input volume. Finally, feature map is flattened for getting a feature vector, which is connected to fully connection layer. This layer uses a softmax activation function [15] to estimate the class membership likelihood (stage1, stage2, stage3, stage4), which corresponds to the infection levels by black Sigatoka of banana plant.

Vision Transformer. It is a transformer designed for computer vision tasks, where the inputs are images divided into a series of patches, and then serialized each one into a vector [8]. These vectors are mapped to smaller dimension and processed by a transformer encoder. Although this scheme has been designed previously for natural language processing tasks, they have shown to be powerful in capturing spatial dependencies in visual data. In details, the proposed ViT receives as input 2D images, which are reshaped from $x \in \mathbb{R}^{H \times W \times C}$ to sequence of flattened 2D patches $x_p \in \mathbb{R}^{C \times N}$ where (H, W) corresponds to the resolution of the original image, C is the number of channels; and finally, N corresponds to the number of patches obtained by $(H \times W)/P^2$, where P is the resolution of each image patch. The output of this process is referred as Patch embeddings. The Position embeddings are added to these patch embeddings to retain positional information (see Eq. 1)

$$\omega_0 = [I_{class}|I_p^1 E; I_p^2 E; \dots; I_p^N E] + E_{pos}, \quad (1)$$

where I corresponds to the image, p is the resolution of each image patch, $E \in \mathbb{R}^{C \times D}$ and $E_{pos} \in \mathbb{R}^{(N+1) \times D}$.

The output of ω is used as input to the Transformer encoder. According to mentioned in [25], the encoder consists of alternating MultiHead self-attention (MSA) layers and MultiLayer perceptron (MLP) blocks. Likewise, a LayerNorm (LN) [3] is applied before each block (see Eqs. 2 and 3). Finally, a linear transformation is applied to obtain the output of proposed model (see Eq. 4).

$$\lambda = MSA(LN(\omega_0)) + \omega_0, \quad (2)$$

$$\beta = MLP(LN(\lambda)) + \lambda \quad (3)$$

$$y = Lin(LN(\beta)) \quad (4)$$

where Lin is used to apply a linear transformation of obtained result of the blocks (MSA) and (MLP). The proposed architecture considering Vision Transformer scheme is showed in Fig. 4.

2.4 Metrics

A set of metrics such as accuracy, precision, recall/sensitivity and F1 score are used to measure the performance of the proposed approach, including a confusion matrix. These metrics are commonly used in classification problems. The obtained values near to 1 suggest more accurate of model. These metrics are formulated as:

$$accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (5)$$

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

$$recall(or)sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$F1Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (8)$$

where TP (true positive), TN (true negative), FP (false positive), FN (false negative).

- TP: model predicts true, and ground-truth is true.
- TN: model predicts false, and ground-truth is false.
- FP: model predicts true, and ground-truth is false.
- FN: model predicts false, and ground-truth is true.

Table 1. Parameter settings of the proposed model (ViTSigat)

Embedding Dims	Layers	Heads	Patches	Patches Size	MLP Size
126	3	3	196	16	256

Table 2. Performance evaluation of the proposed models (CNN and ViTSigat models)

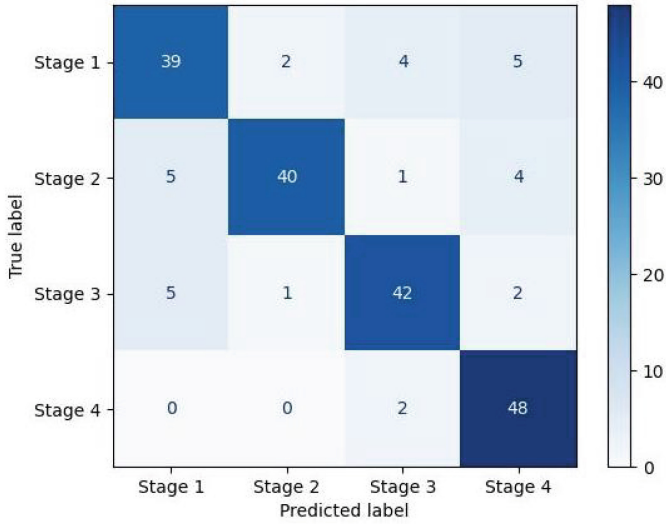
Models	Metric	Stage 1	Stage 2	Stage 3	Stage 4
CNN	Accuracy	0.85			
	Precision	0.78	0.80	0.84	0.96
	Recall	0.80	0.93	0.86	0.81
	F1 Score	0.79	0.86	0.85	0.88
ViTSigat	Accuracy	0.96			
	Precision	0.92	0.94	0.96	1
	Recall	0.96	0.96	1	0.91
	F1 Score	0.94	0.95	0.98	0.95

3 Experimental Results

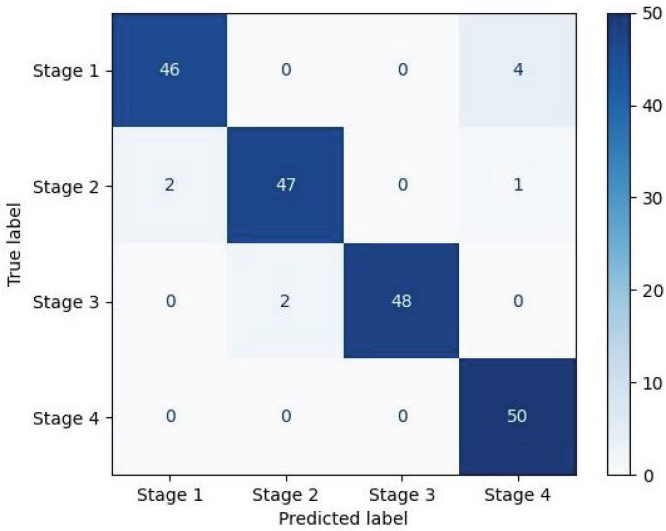
As mentioned above, a ViT model known as ViTSigat is used to detect the disease of black Sigatoka in early stage in leaves banana plants. Additionally, other architecture using CNN has also been considered to tackle this problem. PyTorch library in python is used to implement these models. The details on the experimental results using the proposed models mentioned above are presented in this section. The dataset considered for training of proposed models was presented en Sect. 2.1, which is divided into four stages, each stage had around of 366 RGB images. The images pre-processing were presented in 2.2. Basically, the images were random cropped and resized to 224×224 pixels, and then normalized them considering the mean values of pixels and standard deviation. For training process, a set of 1244 RGB images were used to feed the proposed models, which were trained until 200 epochs. In the evaluation a set of 200 RGB images were considered.

The CNN proposed model in this study was set using two convolutional layers with output channels (32 and 64 respectively), and kernel of 3. MaxPooling 2D is considered with kernel of 2. ReLU as activation function was used for each Convolutional layer and Maxpooling, including a linear transformation as final layer of model. CrossEntropy loss function and Adam optimizer are used to train the network with a learning rate of 0.001, and batch size of 16. The second proposed model (ViT) uses the same optimizer, loss function and learning rate that CNN model mentioned above. Additional variables for training process such as numbers of layers, head, patches of images are set according to the Table 1.

Experimental results obtained with the proposed models are presented in Table 2. The models were trained using different parameters setting. The results



(a)



(b)

Fig. 5. Confusion Matrix of results obtained by proposed models. (a) CNN model (b) ViTSigat model.

of both models were compared between them by using metrics mentioned in Sect. 2.4, being the results of ViTSigat model better than the obtained results by CNN model. Figure 5 shows the confusion matrix of the both models, considering the metrics mentioned previously. The ROC curves were used to compare the

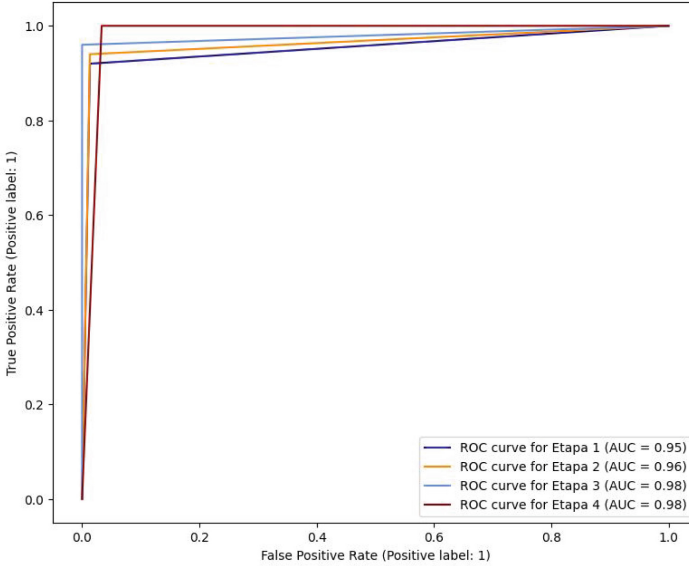


Fig. 6. ROC curves for the proposed model (ViTSigat).

true-positive rate and the false-positive rate of early stage of black Sigatoka for ViTSigat model (see Fig. 6).

3.1 Visualising Relevant Patches

Figure 7 shows example of saliency maps produced by ViTSigat. According to [21], the saliency map is the magnitude of the gradient of the loss function with respect to the pixel intensities, which is used as an indicator to determine if the leaves of banana plants are diseased due to the infection by black Sigatoka. ViTSigat model divides the images in different patches, which allows to focus in relevant areas where the leaves could be affected, and thus, pay more attention to these patches and extract features important for training process. However, a surprising result is that ViTSigat is also sensitive to excessive light changes, which could affect the final result, being essential to capture images correctly.

4 Discussion

As mentioned previously, the obtained results for each proposed model (CNN and ViTSigat) are showed in Table 2. The results of prediction using ViTSigat model show better performance in all metrics (accuracy, precision, recall and F1-Score) if they are compared with CNN proposed model. Likewise, the trainable parameters of proposed models, which allow to evaluate the computational cost,

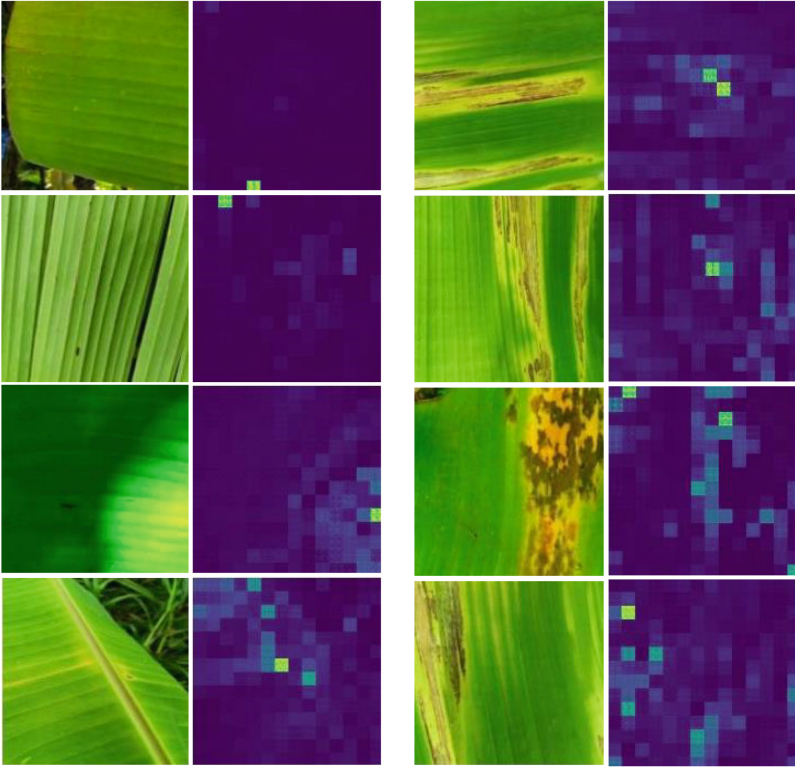


Fig. 7. Saliency maps. This figure shows the relevant patches to identify the early stages of black Sigatoka by using ViTSigat model.

were of 25 M trainable parameters for CNN models and 438K trainable parameters for ViTSigat, being the second model more lighter than the first model. Since the CNN model evaluates all pixels of the image, including non-useful pixels to detect areas where the leaves have been affected by black Sigatoka, the obtained results of this model were less accurate. For the Stage 1, the results of precision and recall metrics (0.78 and 0.80) show that the model is not able of getting relevant features of the leaves of banana plants. This may be due to that the black Sigatoka in Stage 1 are not easily visible since that the leaves of banana plants do not have a significant change in its color, doing that the image have poor texture/ relevant features to be analyzed. However, the obtained results of CNN model for the Stage 4 were better in precision and recall metrics (0.96 and 0.81) than other stages. This is due to the variability of intensity of pixels in the image, since the level of infection of the leaves by black Sigatoka (Stage 4) have rich textures/relevant features, doing visible the significant change in color (green color to yellow/brown/black colors).

Unlike of CNN model, ViTSigat model, which use transformer encoder, are able to learn large order dependence in data sequence, i.e., patches of the images, which include the position embedding of each patch as input data to the transformer encoder. The obtained results were better than CNN proposed model mentioned above. In contrast to the results obtained by CNN model in Stage 1, the obtained results by using ViTSigat model for precision and recall metrics (0.92 and 0.96) are better. This is due to that the model finds relationship/dependence between patches of images, which help to the model to get better results. Furthermore, the analysis of the image for black Sigatoka detection is performed for each patch, whose results are used to feed by transformer encoder. This allowed to focus on important features of input data instead of take attention to non-useful information. Likewise, the obtained results for the Stage 4 for precision and recall metrics (1.0 and 0.91) were better if they are compared with the obtained results for CNN model. This is due to the rich textures of the image when the levels of infection by black Sigatoka are high as well as the usage of patches to divide the image, and thus, analyze each patch of the leaves of banana plants to find significant change in its color.

5 Conclusions

This paper addresses the challenging problem of black Sigatoka detection in early stage in banana plants. This disease can reduce the quality of the product or even lead to total loss of banana production, mainly when the human control is deficient. The preventive control of banana farming has become in a important step to avoid losses of production. During last years, the usage of technological devices such as drones, which can be used to fumigate plantations. However, to reduce cost, it is important to first recognize the disease in the leaves of banana plants before of spraying the whole plantation. In this study, the analysis of black Sigatoka have allowed to divide it in four stage (from 1 to 4) where stage 1 corresponds when the infection in the leaves of banana plants by black Sigatoka are not easily visible, and stage 4 corresponds to an advanced infection by black Sigatoka in the leaves of banana plants, which is visible due to the change in its color. The two proposed models are compared to determine the accuracy of trained models (CNN and ViTSigat models), whose obtained results showed a better performance by using the second model. The manuscript shows that the attention modules of transformer encoder can help to find relevant features and estimate black Sigatoka in early stage more accurately. The obtained precision for black Sigatoka detection using Vision Transformer could be used like a base to develop new proposals that allow take advantage of attention modules, patches and position embeddings. Future works will be focused on extending the usage of multi-spectral images as additional information to the model as well as design hybrid models that allow to use CNN embedded models to filter important features of other inputs data.









References

1. Amara, J., Bouaziz, B., Algergawy, A.: A deep learning-based approach for banana leaf diseases classification. In: *Datenbanksysteme für Business, Technologie und Web* (2017)
2. Anasta, N., Setyawan, F.X.A., Fitriawan, H.: Disease detection in banana trees using an image processing-based thermal camera. *IOP Conf. Ser. Earth Environ. Sci.* **739**(1), 012088 (2021). <https://doi.org/10.1088/1755-1315/739/1/012088>
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
6. Camargo, A., Smith, J.: Image pattern classification for the identification of disease causing agents in plants. *Comput. Electron. Agric.* **66**(2), 121–125 (2009). <https://doi.org/10.1016/j.compag.2009.01.003>
7. Continuous Agricultural Surface and Production Survey (ESPAC): Technical bulletin. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_agropecuarias/espac/2023/Boletin_tecnico_ESPAC_2023.pdf (2024). Accessed 7 Jun 2024
8. Dosovitskiy, A., et al.: An image is worth 16 x 16 words: transformers for image recognition at scale. *ArXiv abs/2010.11929* (2020)
9. Food and Agriculture Organization of the United Nations (FAO): Banana market analysis. <https://openknowledge.fao.org/server/api/core/bitstreams/b7c671a6-c669-4f54-a7d2-c2fc9e4a489f/content> (2023). Accessed 7 Jun 2024
10. Geurts Pierre, E.D., Wehenkel, L.: Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006). <https://doi.org/10.1007/s10994-006-6226-1>
11. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
12. Keerthi, S.S., Chapelle, O., DeCoste, D.: Building support vector machines with reduced classifier complexity. *J. Mach. Learn. Res.* **7**(55), 1493–1515 (2006)
13. Li, R.H., Belford, G.G.: Instability of decision tree classification algorithms. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 570–575 (2002). <https://doi.org/10.1145/775047.775131>
14. Liao, W., Ochoa, D., Gao, L., Zhang, B., Philips, W.: Morphological analysis for banana disease detection in close range hyperspectral remote sensing images. In: *IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3697–3700 (2019). <https://doi.org/10.1109/IGARSS.2019.8899087>
15. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295* (2016)
16. Ma, Z., Kaban, A.: K-nearest-neighbours with a novel similarity measure for intrusion detection. In: *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pp. 266–271 (2013). <https://doi.org/10.1109/UKCI.2013.6651315>
17. Mortensen, E.N., Deng, H., Shapiro, L.: A sift descriptor with global context. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 184–190. IEEE (2005)
18. Owomugisha, G., Quinn, J.A., Mwebaze, E., Lwasa, J.: Automated vision-based diagnosis of banana bacterial wilt disease and black sigatoka disease. <https://api.semanticscholar.org/CorpusID:32978365>

19. Samridhi, S., Kalpana, M., Parimalarangan, R., Palanichamy, N.V.: Identification of Sigatoka leaf spot disease in banana using convolutional neural network (CNN). *Asian J. Agric. Exten. Econ. Sociol.* **41**(9), 931–936 (2023)
20. Senthil Kumar, A., Ademe, M., Ananda Kumar, K.S., Adusumalli, S., Venkata Subbarao, M., Sudhakar, K.: Detection of leaf black sigatoka disease in enset using convolutional neural network. In: Zen, H., Dasari, N.M., Latha, Y.M., Rao, S.S. (eds.) *Soft Computing and Signal Processing*, pp. 301–310. Springer Nature Singapore, Singapore (2024)
21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013)
22. Singh, P., Kumar, A., Shobhit, Chandel, R., Dongre, P.: Deep learning insights into banana sigatoka disease: Resnext50 for seriousness classification. In: *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2, pp. 1–6 (2024). <https://doi.org/10.1109/IATMSI60426.2024.10503074>
23. Singh, V., Misra, A.: Detection of plant leaf diseases using image segmentation and soft computing techniques. *Inf. Process. Agric.* **4**(1), 41–49 (2017). <https://doi.org/10.1016/j.inpa.2016.10.005>
24. Sneddon, I.N.: *Fourier transforms*. Courier Corporation (1995)
25. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017)
26. Vidhya, N., Priya, R.: Detection and classification of banana leaf diseases using machine learning and deep learning algorithms. In: *2022 IEEE 19th India Council International Conference (INDICON)*, pp. 1–6 (2022). <https://doi.org/10.1109/INDICON56171.2022.10039912>
27. Zhang, D., Zhang, D.: Wavelet transform. In: *Fundamentals of image data mining: analysis, features, classification and retrieval*, pp. 35–44 (2019)



Automatic Parking Space Segmentation Using K-Means Clustering and Image Processing Techniques

Anthony Xavier Romero Gonzalez , Kevin Sebastian Campoverde Ambrosi ,
Patricio Eduardo Ramon Celi , Alexandra Bermeo , Marcos Orellana ,
Jorge Luis Zambrano-Martinez  , and Patricio Santiago García-Montero 

Computer Science Research & Development Laboratory (LIDI), Universidad del
Azuay, Cuenca, Ecuador

{anthony.romero,kevin.campoverde,patricio.ramon,
santyg20}@es.uazuay.edu.ec,
{alexbermeo,marore,jorge.zambrano}@uazuay.edu.ec

Abstract. Proper management of parking spaces is essential in urban environments. This study proposes an approach for parking space segmentation using the K-means algorithm and the OpenCV library. The main objective is to determine the trapezoid describing the parking area by analyzing data previously collected from multiple photographs. These images contain several vehicles parked in different dispositions and moments in time. For this, the coordinates of the four leading edges that compose each car were considered. The previously obtained data were used to estimate the trapezoid defining each photograph's parking zone. This approach combines segmentation and image processing techniques to delimit parking spaces in urban environments.

Keywords: DBSCAN · K-means · OpenCV · Segmentation of smart parking · YOLO

1 Introduction

The detection of free spaces concerning parking occupancy is a challenge in parking applications or smart sites. In this context, using deep neural networks is effective and optimal for solving this problem. Carrasco et al. [5] approach vehicle detection in parking lots using the You Only Look Once (YOLO) model and convolutional neural networks at multiple scales. Their version is specifically designed to detect small cars from a zenith view. This modification uses a multi-scale module and spatial/channel attention mechanisms to improve performance in this application. An additional advantage of the proposed modification is its ability to detect small vehicles with high accuracy, outperforming even the multi-version solution of the YOLO-v5 model. It was concluded that the modification decreased detection speed compared to the small and medium car profiles of the YOLO-v5 algorithm [11]. However, it still outperforms the large car profiles.

Therefore, the modified model achieves an effective balance between accuracy and speed. In addition, the results have implications for developing intelligent parking systems.

In their research, Alsheikhy et al. [1] apply a convolutional neural network model to detect parked lines in images. Their approach uses Canny edge detection, Gaussian blur, and Hough EDGE detection techniques. These techniques allow the identification of horizontal and vertical lines that compound the parking grid. In another study by Athira et al. [2], they use the “OCR-based parking slot detection system” model. The implementation consists of assigning numbers to each parking slot, physically painting them, and processing the image using Optical Character Recognition (OCR) to identify the presence of the numbers. In image recognition and label assignment, the authors of [12] propose an approach based on YOLO-v4 to identify objects within an image and classify them using labels. Specifically, vehicle recognition uses a methodology in which a bounding box of four coordinates is drawn around each detected vehicle. The corresponding parking area is established based on the identified vehicles’ position.

In addition, within this project, the proposed solution for a parking lot equipped with an intelligent system is presented to achieve an optimal use of space. This project will benefit the community by allowing them to quickly find a suitable parking space, especially during peak hours. It has been observed that particular lot A saturates faster than parking lot B. By implementing this multi-parking comparison system, users can select the location with available spaces, thus avoiding the frustration of visiting busy parking lots and optimizing their search time.

Parking space recognition technology is based on identifying the type of space and determining the parking space’s entrance and direction. Determining the type of parking space is mainly done by placing the corner points of the space [18]. The localization of detected objects and their object recognition/classification are integral parts of the machine vision-based object detection and recognition algorithm. Machine vision-based systems using image processing techniques fall into recognition-based and appearance-based categories [6]. By learning the features of cars and classifying them from the input image regions, the recognition-based approach aims to arrange them into predefined parking spaces. In contrast, appearance-based calculations calculate space from appearance features such as orientation histogram density, masked area, or image subtraction.

The detection of parking stalls using YOLO-v3 uses a Darknet structure and two DBL units used repeatedly in the YOLO-v3 architecture to form residual blocks and build a deep and efficient network, mitigating problems such as gradient degradation [8]. Moreover, it applies convolutions and cascade operations to extract features from the input image and two ResNet units [10]. Its architecture comprises residual blocks that can be stacked to form networks without slowing down performance. Clustering with K-means is also included, which classifies the parking stall space [8]. The theoretical framework focuses on parking efficiency, using occupancy and duration as critical indicators. A prediction system consisting of two modules is proposed: occupancy prediction and duration estimation.

The K-means clustering algorithm is used to classify parking spaces with similar patterns. An occupancy prediction model is trained for each group, while regression models the duration. Combining both indicators provides drivers with information about empty parking spaces [15]. Moreover, in the study presented by [14], the authors use the path-finding algorithm-based Ant Colony Optimization (ACO) approach, whose aim is to find an acceptable path with graphs and heuristic knowledge to find optimal solutions in a parking lot study.

Once the object is identified, the object metric, built based on the YOLO-v3 structure, is obtained, and the coordinates of the four vertices of the quadrilateral accurately encompassing the parking space are estimated. Standard object detectors, such as YOLO-v3, calculate the width, height, and center coordinates since their goal is to find a horizontal bounding box enclosing an object [9]. In this context, the technique of comparing the vertices of different bounding boxes generated based on the dimensions of several cars in the same area is employed.

This technique forms clusters corresponding to each vertex of different bounding boxes within an area. By obtaining the cluster centers of each vertex, relevant information can be generated to establish parking spaces based on the data obtained from multiple cars. They are assigning an appropriate space to make known the spaces where vehicles can be parked.

Finally, this document is organized as follows: Sect. 2 contains the methodology of the work. Section 3 presents the results, and finally, Sect. 4 contains the conclusions of this research.

2 Methodology

The methodology presented in Fig. 1 consists of four stages: i) Data preprocessing, ii) Delimitation of possible clusters, iii) Clustering application, and iv) Space segmentation. In each stage, the input and output files are included along with the implemented guidelines. The first stage reduces rows with null values and reduces data noise. The delimitation of possible clusters determines the maximum number of clusters with K-means to analyze their Silhouette coefficient. The next stage, the clustering application, generates a dataset with the optimal number of clusters and their label. Finally, in the space segmentation, the parking spaces are generated in the analysis image under study by applying Open Source Computer Vision (OpenCV) algorithms.

2.1 Data Preprocessing

This activity discusses the data set and the cleaning techniques used. The data consists of corner coordinate information of rectangles representing parked cars in images. The images belong to the dataset presented in the work of Almeida et al. [7]; the dataset contains 695,899 images of two parking lots in different scenarios and environmental variations, which makes it a robust dataset for research purposes. In this case, 900 photos of one of the parking lots have been used to detect the vehicles present using YOLOv8 and generate the dataset with the coordinates of the vehicles.

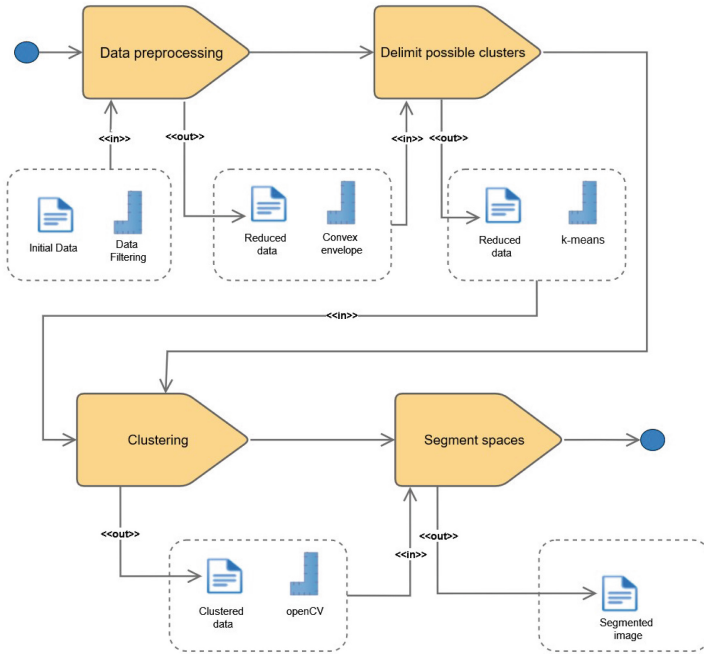


Fig. 1. Research Methodology Phases represented with SPEM 2.0

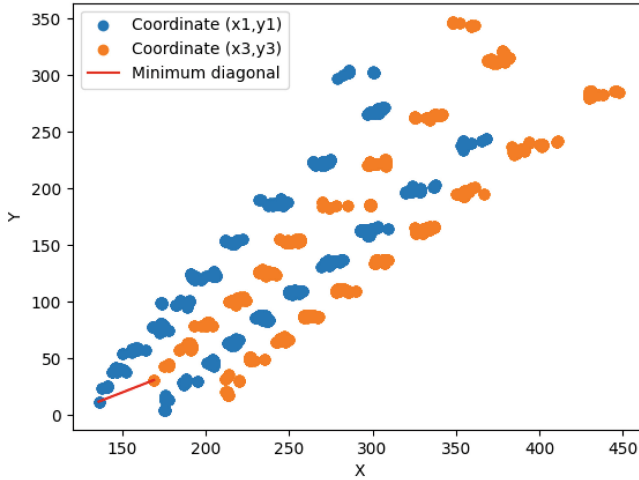
For the detection, the pre-trained model of greater precision YOLOv8x was used with the predicted mode to run the predictions in the image directory. The *classes* parameter was also configured to detect only the vehicles in the images. The 'Cluster' field is included with integer values corresponding to the previously calculated cluster number. Likewise, each cluster's x and y coordinates of the centroid are included. Since the features of interest comprise the corner coordinates, only consider the columns corresponding to points in the x and y axes. The initial dataset features are presented in Table 1. The filtering process involves the removal of null data, along with the application of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN is a density-based clustering algorithm that is particularly effective for identifying outliers and noise in the data. It helps to remove coordinates far from the center points. The parameters used in DBSCAN include an epsilon value of 0.5 and a minimum number of samples set at five. These values represent the default parameters of the algorithm. They are considered optimal in terms of performance and representation of the underlying structure of the data [17].

2.2 The Delimitation Number of K-Means

The second stage of the study seeks to determine the number of potential clusters and calculate their silhouette coefficient. Therefore, it is necessary to analyze the

Table 1. Dataset characteristics

Features	Description
Type	Cluster of vehicles in a parking lot
Size	10,829 records
Coordinates	4 for each parking lot
Maximum number of clusters	21

**Fig. 2.** Minimum diagonal between x_1 and x_3 coordinates

complete set of coordinates that compound the parking space, identifying the value of the smallest diagonal that spans from the point (x_1, y_1) to the end (x_3, y_3) , as shown in Fig. 2.

In the study, the convex envelope is calculated to determine the value of the area of the points that compound the parking lot. This structure is fundamental in computational geometry and consists of the most miniature set of points that completely encloses a given set of points in an Euclidean space [13]. By joining these points, a polygon is obtained. The generated polygons' area represents the total area of the parking lots, as shown in Fig. 3. The ratio between the polygon's total area and the minor diagonals' rectangular area determines the maximum possible number of parking lots, obtaining a value of 88.

2.3 Clustering K-Means

As Bishop [3] described, the K-means algorithm is a powerful clustering method used to identify groups in data sets. It is particularly effective in partitioning data into a predefined number of clusters to minimize the distance within each

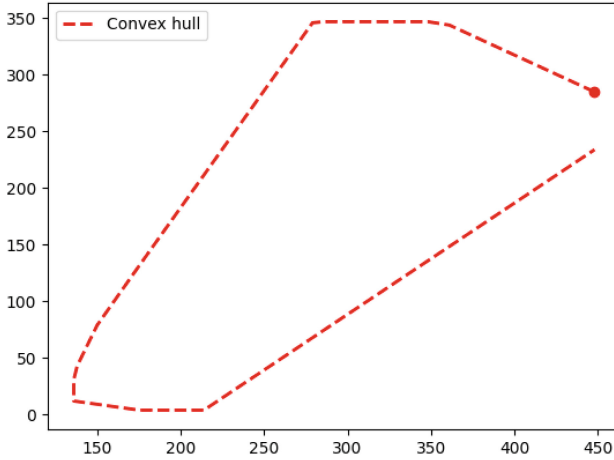


Fig. 3. Convex Envelope

cluster and maximize the distance between clusters. In our study, we chose the K-means algorithm for its unique ability to cluster the edge coordinate data of parked vehicles in images, enabling accurate segmentation and delimitation of parking spaces.

The silhouette index is a metric that quantifies the cohesion and separation of the groups obtained by the clustering algorithm [16]. Properly managing parking spaces in urban environments is necessary to ensure accuracy and reliable segmentation. Using the silhouette index allowed us to quantify the cohesion and separation of the clusters formed by the algorithm, which presents valuable information about the quality of the segmentation performed. A high silhouette index value indicates that the data within each cluster is highly cohesive and well separated from the data in other clusters [16].

Figure 4 shows the silhouette coefficient as a function of the number of clusters. A prominent peak in the silhouette coefficient value was identified. This peak indicates the presence of an optimal structure in the data, suggesting an adequate number of clusters. Therefore, the optimal number of clusters for parking space segmentation was determined to be 21, with a coefficient of 0.8527.

The proposed approach has proven its effectiveness in accurately segmenting parking spaces in urban environments. Figure 5 presents the identified parking spaces with noise, while Fig. 6 presents the analysis performed on data without noise, confirming the proposal's validity and reliability.

2.4 Modifying the Image to Plot Rectangles

The cv2 module, the main module in OpenCV Library [4], was used to display the results. This library has several functionalities; in this case, it is used for reading

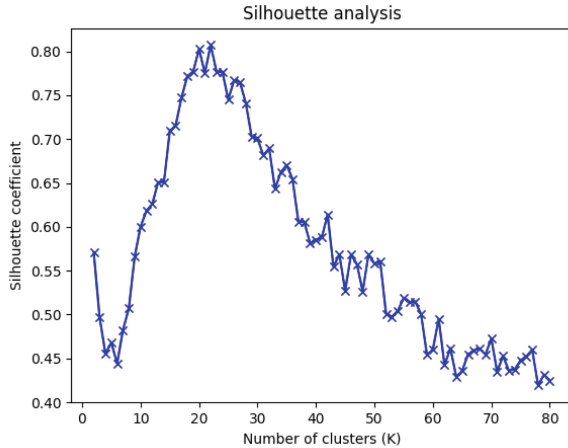


Fig. 4. Relationship Between Silhouette Coefficient and Number of Clusters

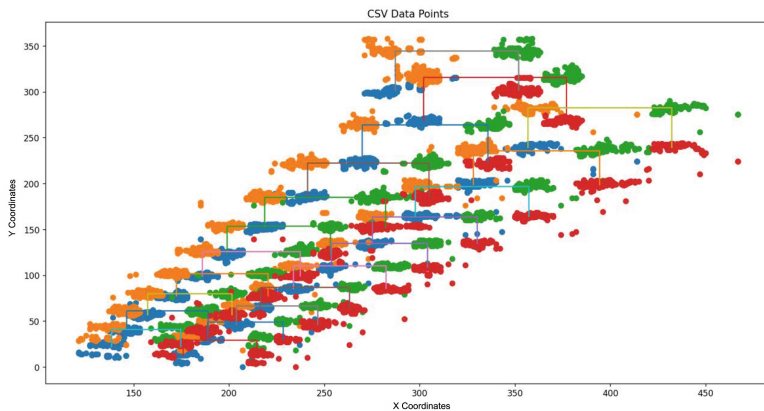


Fig. 5. Two-Dimensional Visualization of Clustered Data with Noise

and writing images. The image of the empty parking lot had to be modified to display the data because the coordinates generated with the previously processed data do not coincide with the parking spaces, as shown in Fig. 7.

A process needed to be performed to draw the parking spaces. The required section from the image was cropped, then resized to 640 pixels wide and 360 pixels high, and finally, the original image of the parking lot was tilted. This is so that the coordinates of the clusters match the parking spaces within the image. The characteristics presented in Fig. 8 were considered to define the changes made in the image.

There are some remarkable differences when comparing the original image of the parking lot in Fig. 7 and the example of how the data should be displayed

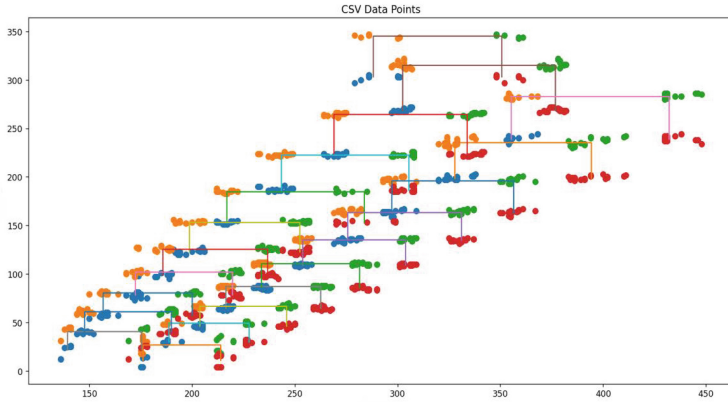


Fig. 6. Noise-Filtered Visualization of Clusters in Two Dimensions



Fig. 7. Empty original image of parking lot

in Fig. 8. The difference can be minimized by modifying the image parameters mentioned above. After adjusting the image, the parking spaces are drawn in a way that corresponds to the image.

The rectangles were established through the function “rectangle,” contained in the cv2 library. The parameters of this function receive the image (modified) in two coordinates: the upper left corner of the rectangle and the lower right corner. Finally, it gets the color of the lines in Red, Green, Blue (RGB) format with numeric values and the desired width of the line.

For all the rectangles corresponding to the parking spaces in the plot, a loop with the function “rectangle” was used, changing only the parameter of the coordinates obtained from each cluster. Red was used for the color, which in RGB is represented by (0, 0, 255) with a width of 2.

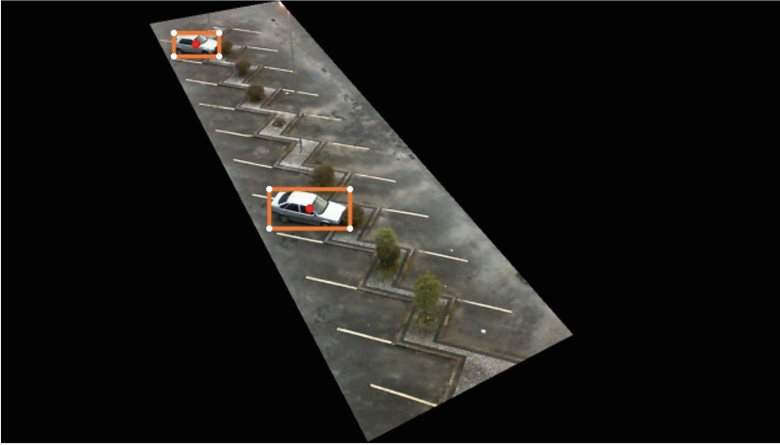


Fig. 8. Coordinates and centroid information plotted on the image

3 Results

Figure 9 presents the visual results of the 21 parking spaces identified using only the k-means method for clustering. The presence of noise in the data significantly affects the generation of parking spaces. By not eliminating this data, an overlap of spaces occurs, making their identification considerably more difficult. These results visually represent the distribution of the segmented parking spaces in the study area (the convex enclosure).

In contrast, Fig. 10 shows the 21 segmented spaces using an alternative method, which involves DBSCAN to refine the data, identifying and eliminating noise data, and then employing k-means for data clustering. Upon comparison, the second method was chosen for its ability to generate accurate rectangles. It provides a clear understanding of the available spaces within the image, reassuring the audience about its effectiveness.

The image provides insight into the distribution of parking spaces in the analyzed area and how they are grouped according to similar characteristics. This information can be helpful for efficient parking planning and management and identify patterns and trends related to parking spaces in each area. It is important to note that the clustering algorithms (DBSCAN and K-Means) were used with the coordinates of the vehicles detected in the images only to generate the parking spaces, not to predict new coordinates.

Also, it is essential to highlight the capacity of vehicle identification so that the *YOLOv8x* mod has a mean accuracy (mAP) of 100% in the 21 test parking spots. Although the most demanding model regarding computational resources, it achieves remarkable results even in the most challenging locations, such as the upper corners, where vehicles are often visually obstructed due to the camera's position.



Fig. 9. Parking spaces plotted with unfiltered noise data



Fig. 10. Parking spaces plotted with filtered noise data

4 Conclusion

The DBSCAN default parameters effectively improved the quality of the coordinates by eliminating isolated points, resulting in more accurate results. In addition, the maximum possible number of parking lots was calculated by limiting the total area using the convex enclosure, significantly reducing the iteration time.

This paper presents a methodology for obtaining accurate information about parking spaces. The k-means result is effective in segmenting such spaces; it is

essential to note that silhouette analysis is of utmost importance for this type of problem to improve the accuracy of the k-means algorithm.

Within the graph, the assembly of rectangles is a complex task, influenced by the perspective of the image and the height parameter of the cars. This complexity underscores the need for a sophisticated solution. However, the result we have obtained meets the challenge of segmenting the parking spaces of a parking lot by grouping the study data.

Future work is planned to use this methodology in the parking lot of the University of Azuay and optimize the algorithm for the segmentation of spaces in real-time; it is expected to significantly improve parking management, speeding up the search for spaces and reducing congestion. In addition, it is planned to explore the integration of this methodology in Internet of Things (IoT) devices to optimize processes, reduce energy consumption, and increase efficiency.

Acknowledgments. This work was supported by the vice rectorate of Research at Universidad del Azuay. Therefore, we thank them for their financial and academic support and the entire Computer Science Research & Development Laboratory (LIDI) staff.

References

1. Alsheikhy, A.A., Shawly, T., Said, Y.F., Lahza, H.: An intelligent smart parking system using convolutional neural network. *J. Sensors* **2022**(1), 7571716 (2022)
2. Athira, A., Lekshmi, S., Vijayan, P., Kurian, B.: Smart parking system based on optical character recognition. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1184–1188. IEEE (2019)
3. Bishop, C.M., Nasrabadi, N.M.: *Pattern Recognition and Machine Learning*, vol. 4. Springer (2006)
4. Bradski, G., Kaehler, A.: *Learning OpenCV: Computer Vision with the OpenCV library*. O'Reilly Media, Inc. (2008)
5. Carrasco, D.P., Rashwan, H.A., García, M.Á., Puig, D.: T-yolo: Tiny vehicle detection based on yolo and multi-scale convolutional neural networks. *IEEE Access* **11**, 22430–22440 (2021)
6. Choeychuen, K.: Automatic parking lot mapping for available parking space detection. In: 2013 5th International Conference on Knowledge and Smart Technology (KST), pp. 117–121. IEEE (2013)
7. de Almeida, P.R., Oliveira, L.S., Britto, A.S., Silva, E.J., Koerich, A.L.: Pklot—a robust dataset for parking lot classification. *Expert Syst. Appl.* **42**(11), 4937–4949 (2015)
8. Ding, X., Yang, R.: Vehicle and parking space detection based on improved yolo network model. *J. Phys. Conf. Ser.* **1325**, 012084 (2019)
9. Do, H., Choi, J.Y.: Context-based parking slot detection with a realistic dataset. *IEEE Access* **8**, 171551–171559 (2020)
10. Fang, J., Wang, P.: Application of improved YOLO v3 algorithm for target detection in echo image of sonar under reverb. *J. Phys. Conf. Ser.* **1748**, 042048 (2021)
11. Jocher, G., et al.: *Ultralytics/YOLOv5: v3. 1-bug fixes and performance improvements*. Zenodo (2020)

12. Patel, R., Meduri, P.: Car detection based algorithm for automatic parking space detection. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1418–1423. IEEE (2020)
13. Preparata, F.P., Shamos, M.L.: Computational Geometry: An Introduction. Springer Science & Business Media (2012)
14. Shao, W., Salim, F.D., Gu, T., Dinh, N.T., Chan, J.: Traveling officer problem: managing car parking violations efficiently using sensor data. *IEEE Internet Things J.* **5**(2), 802–810 (2017)
15. Shao, W., Zhang, Y., Guo, B., Qin, K., Chan, J., Salim, F.D.: Parking availability prediction with long short term memory model. In: Green, Pervasive, and Cloud Computing: 13th International Conference, GPC 2018, Hangzhou, China, May 11–13, 2018, Revised Selected Papers 13, pp. 124–137. Springer (2019)
16. Starczewski, A., Krzyżak, A.: Performance evaluation of the silhouette index. In: Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14–18, 2015, Proceedings, Part II 14, pp. 49–58. Springer (2015)
17. Stenborg, T., Silversides, K.: Low-cost Paretonian DBSCAN parameter estimation for Sklearn. In: Australian Data Science Network I. ADSN Conference Series, vol. 1, p. 8 (2022)
18. Zhang, J., et al.: An improved parking space recognition algorithm based on panoramic vision. *Multimedia Tools Appl.* **80**, 18181–18209 (2021)

Artificial Intelligence and Machine Learning Applications



Neural Agents with Continual Learning Capacities

Luis Zhinin-Vera^{1,2,3}(✉) , Elena Pretel¹ , Alejandro Moya¹ ,
Javier Jiménez-Ruescas¹ , and Jaime Astudillo² 

¹ LoUISE Research Group, University of Castilla-La Mancha, 02071 Albacete, Spain
luis.zhinin@uclm.es

² School of Mathematical and Computational Sciences, Yachay Tech University,
100650 Urcuquí, Ecuador

³ MIND Research Group—Model Intelligent Networks Development, Urcuquí,
Ecuador

Abstract. The contemporary Artificial Neural Networks (*ANNs*) often suffer from catastrophic forgetting, where learned parameters are overwritten by new tasks. This paper presents a novel approach using a Reinforcement Learning (RL) agent with Continual Learning (CL) capabilities to navigate a visual robotic structure, achieving advanced proficiency in Tic-Tac-Toe. The system integrates a webcam for environmental perception, specialized neural blocks for feature extraction, and a communication bus linking self-taught agents with advisors. A knowledge protection mechanism prevents the loss of acquired parameters during new learning iterations. The methodology was validated on a physical robot, implemented with C++ and OpenCV, demonstrating its ability to retain knowledge and enhance gameplay, effectively emulating intelligent children’s learning strategies. The proposed system was tested in a real-world setting, achieving an average accuracy of 92% in task completion and demonstrating a 15% improvement in task retention over traditional methods.

Keywords: continual learning · neural agents · reinforcement learning

1 Introduction

Continual Learning (*CL*) is crucial for intelligence. Humans have the innate ability to learn new tasks without forgetting previous knowledge. Notably, knowledge acquired at one stage or age is used to learn more complex rules in subsequent learning events. In contrast, Artificial Neural Networks (*ANN*’s) suffer from *catastrophic forgetting*, where new learning processes overwrite previously learned information. It remains an open scientific question how knowledge from one stage supports the acquisition of new abilities. An illustrative example involves training a neural network for task *A* by tuning hyperparameters efficiently. However, incorporating an additional task *B* using gradient descent

algorithms leads to catastrophic forgetting, eroding the learning from task A [14, 24]. A potential solution is to train for both tasks simultaneously, creating a single model, but this approach has significant drawbacks and contradicts biological principles of sequential learning.

In animal learning, knowledge acquisition occurs through a sequence of learning peaks from infancy to adulthood. Learning individuals exhibit an internal drive that fluctuates with age and complexity. During early learning stages, random decision-making is prevalent, akin to a reinforcement learning (RL) agent exploring new environments for rewards and developing local policies.

Biological agents appear to store the resultant neural network and synaptic structures from learning processes securely, preserving valuable policy information for future use. Recent experiments suggest that the mammalian brain avoids catastrophic forgetting by protecting previously acquired knowledge in neocortical circuits [11, 23]. However, the exact mechanisms of this preservation remain unclear. Additionally, the stored policy is ready to support new learning processes, facilitating continual learning.

The RL algorithm proposed in this paper incorporates these biological principles by focusing on: agent self-sufficiency consolidating and safeguarding the *net-weights structures* representing learned policies and ensuring the learned policy supports and enhances future learning processes. The primary objectives of this study are to develop a neural agent with continual learning capacities and to demonstrate its application in a real-world task, specifically the game of Tic-Tac-Toe. The key contributions of this work include (1) the design of a dynamic reinforcement learning framework that mitigates catastrophic forgetting, (2) the implementation of a knowledge protection mechanism within the neural architecture, and (3) the successful application of this framework in a physical robotic system, showcasing its effectiveness in continual learning scenarios.

Previous works have applied neural self-motivation concepts in robot controllers using biologically inspired models [7–9]. These robots learned high-level Tic-tac-toe playing policies autonomously and competed effectively with humans. This work is structured as follows. Section 2 reviews related works concerning the proposed method. Section 3 shows an important background about the definitions. Section 4 presents the methods and the proposed algorithm. Section 5 shows the results of multiple experiments performed. Finally, Sect. 6 outlines conclusions and future work.

2 Related Works

ANNs require specific training to work effectively once trained [27]. This approach contrasts with natural learning processes observed in humans and animals, which involve continuous and diverse data acquisition [18]. While humans and animals can continuously acquire skills without significant interference with previously learned ones [30], AI-based models often suffer from catastrophic forgetting. This phenomenon describes the degradation in performance of a model trained for one task when subsequently trained for another [24].

Various research efforts address catastrophic forgetting by developing continual learning models capable of adapting to and learning from a continuous stream of information [27]. For instance, Shin et al. [30] proposed the Deep Generative Replay model, inspired by the hippocampus, which combines a deep generative model (generator) with a task-solving mechanism (solver). Lopez-Paz and Ranzato [21] introduced Gradient Episodic Memory (GEM), which allows the transfer of acquired knowledge across tasks, mitigating catastrophic forgetting.

Kompella et al. [16] presented Continual Curiosity-driven Skill Acquisition (CCSA), enabling motivated robots to acquire, store, and reuse skills by associating rewards with compact representations of visual data. Ashfahani and Pratama [1] developed Autonomous Deep Learning (ADL), a flexible deep learning algorithm capable of dynamically adjusting network structures to combat catastrophic forgetting through mechanisms like node growth and pruning.

This paper introduces a neural computational model inspired by biological learning processes, where agents progressively awaken and learn at different stages, embedding knowledge within energy-efficient neuron structures acting as independent agents advised by trainable networks known as advisers. This framework facilitates the development of a system capable of playing Tic-tac-toe at a level comparable to a creative child. Unlike previous studies that focus on static neural network architectures prone to catastrophic forgetting, our approach introduces a dynamic reinforcement learning agent with continual learning capabilities, allowing it to adapt and retain knowledge across different tasks. This approach not only mitigates catastrophic forgetting but also enhances the agent’s ability to apply learned policies to new, more complex tasks, a feature not addressed in the reviewed studies.

The model incorporates an energy barrier within neural agents to safeguard acquired knowledge from catastrophic forgetting. This barrier ensures that valuable knowledge remains untouched and supportive of ongoing learning processes, immune to unintended weight modifications caused by subsequent learning algorithms. Experimental validation involved implementing this approach in a visual-physical robot, demonstrating its ability to autonomously master Tic-tac-toe.

3 Background

3.1 Exploration

Young children are exceptional learners, prompting contemporary AI research to incorporate child-like behavior and exploration strategies into artificial agents [17]. This paper proposes an agent using continual learning to play Tic-tac-toe, mimicking children’s learning processes.

3.2 Tic-Tac-Toe

Tic-tac-toe, an ancient game with records dating back to Egypt [33], is widely used in gameplay algorithm research for its reliable rules and strategic complexity. In our study, Tic-tac-toe serves as a proof of concept for our continual

learning approach. Despite its simplicity, the $3 \times 3 \times 3$ board has nearly $9!$ possible games, offering a vast search space suitable for evaluating adversarial search problems.

Tic-tac-toe has historically been foundational in reinforcement learning research, exemplified by MENACE (Machine Educable Noughts And Crosses Engine) [25], a pioneering work before modern machine learning formalization. Its combinatorial nature also makes it a subject in automata theory and formal languages, alongside Chess and Go. A brute force approach to Tic-tac-toe has a complexity of $\mathcal{O}(n!)$, necessitating clever methods for higher dimensions, such as $4 \times 4 \times 4$ in 3D Tic-tac-toe. The game extends into n dimensions, with $3n^2 + 6n + 4 = 3(n+1)^2 + 1$ winning lines for the $n \times n \times n$ version, making it ideal for studying complete information games.

Comprehensive studies, such as those by J'ozsef [3] and Golomb et al. [15], highlight the game's complexity from combinatorial and probabilistic perspectives. While our study focuses on the $3 \times 3 \times 3$ version, we believe its complexity offers a challenging platform for new learning approaches and future investigations into general cases.

3.3 Rewards

First, the game logic must be framed in reinforcement learning terminology [32]. Typically, a learning agent receives a reward of -1 for placing its token in an occupied zone, 0 for making a valid move, and 0 when the match ends in a tie. The objective is to maximize the reward, such as 100 for winning the game by aligning three tokens on the 2D board.

3.4 An Even Bigger Reward

In positional games like Tic-tac-toe, certain initial moves known as “openings” offer strategic advantages to skilled players. Through play, both children and adults often independently discover openings that guarantee a decisive advantage, ensuring victory within four moves. This forward-looking sequence of moves, predicting victory regardless of the opponent's actions within a short span, is highly valued in more sophisticated games such as Chess and Go, often associated with individuals termed as prodigies.

Cutting-edge AI research focuses on enhancing lookahead capabilities and developing learning agents that emulate human learning behaviors [12]. Notable examples include Google's advancements in Go [6, 20] and cybersecurity systems employing lookahead strategies to detect hackers [17].

The mechanisms by which humans develop lookahead proficiency through continual learning remain under investigation. However, it is widely acknowledged that this complex process involves a blend of long-term and short-term memory mechanisms, internal intelligent agents discovering deep logical connections, and the sequential establishment of communication pathways.

3.5 Continual Learning Layout

In this study, RL agents address continual learning in two stages, enhancing their ability to play Tic-tac-toe until their performance resembles that of clever children. Figure 1 illustrates the overall layout employed in this research.

The robot uses a webcam to observe a Tic-tac-toe board. A convolutional network processes the video, generating sparse codes for a fully connected network, the adviser. The adviser produces a policy indicating where to place the next token; untrained networks generate random values. The adviser’s output influences a neural agent, which generates its own policy and directs the robotic arm to place a token on the board for each new state it perceives.

The convolutional network (b) processes a 480×480 2D color image of the board. Each square is converted to a 31×31 pixel matrix and convolved with four 5×5 filters for diagonal, horizontal, and vertical lines, with a stride of two. The resulting image undergoes max pooling and passes through a fully connected network. It is trained using gradient descent to produce sparse codes in three output neurons: “000” for empty, “100” for (X), and “001” for (O). The combined nine squares yield 27 lines, representing the sparse board coding.

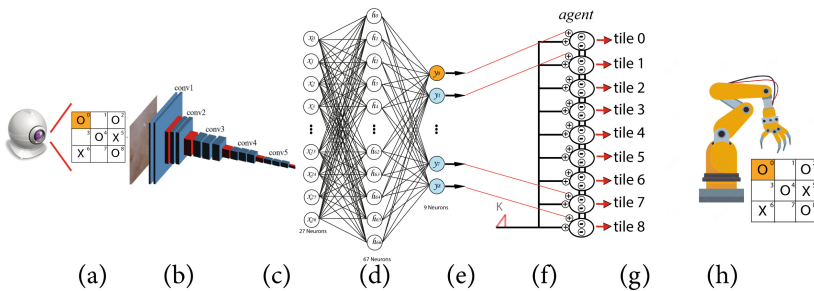


Fig. 1. Continual Learning Layout: (a) A robot observes a physical board via a webcam. (b) A trained convolutional network converts the board image to sparse code (c), which is sent to the adviser (d), a fully connected classifier network that, once trained, suggests possible playing policies. (e) These policies are delivered to a self-motivated neural agent (f), which uses the adviser’s suggestions to generate its policy (g), burning energy in the process. The agent’s policy follows a Markov chain, directing the robot arm (h) to place physical tokens on the board.

3.6 The Biological Connection

As mentioned, young children are exceptional learners. Thus, it seems relevant to explore models of learning agents that mimic the behavior of young children. Following AI research trends, our model agent learns to play high-level Tic-tac-toe in two stages. In the first stage, akin to an impulsive infant, the agent learns *token location*, placing its tokens (O) in empty squares and alternating turns

with the opponent (X). This stage concludes once the agent masters this task. In the second stage, the agent focuses on advanced game strategies, leveraging the previously learned policy and the Bellman equation’s three terms [9].

The agent, resembling biological systems, comprises specialized blocks of sigmoidal neurons. These neurons consume *dark energy* seemingly dissipating without purpose. RL agents navigate new situations, accumulating optimal policies for token placement and strategy. Memory formation occurs through sparse code processing by specialized adviser networks in each stage. Policy knowledge is stored in a read-only network, with weights fixed post-stage completion.

Humans excel in adapting, acquiring, and refining skills, transferring them across domains [2, 4, 31]. Brown and Kane [5] demonstrated that children could transfer principles like “mimicry as a defense mechanism” between animals. Fong et al. [13] showed successful knowledge transfer from statistics to everyday problems. Chen and Klahr [10] found that learning can transfer across different problems and formats.

Humans also excel in classifying stimuli along multiple dimensions. Children can categorize animals by size and later by ferocity or color without losing previous knowledge. This continual learning is due to hippocampal mechanisms, interleaving ongoing experiences with recollected past training samples [19, 22]. Neural networks store and *replay* past memories, avoiding destructive interference and reinforcing relevant representations through interleaving [26, 28, 29].

3.7 The Bellman Equation

The Bellman Equation expresses a relationship between the value of a state and the values of its successor states. Bellman equation [32] is defined as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]. \quad (1)$$

The learned action-value function Q directly approximates the optimal action-value function, independent of the followed policy, and addresses optimal control of differential-difference (time-lag) processes.

The Bellman equation does not specify the magnitude of a dimension. Depending on the simulated environment and system, the time difference required can range from microseconds to years. Typically, the equation is applied over a fixed timespan, such as in a deep convolutional network performing image processing, where backpropagation occurs at maximum speed ($1 \mu\text{s}/\text{core}$). After this stage, training ceases, the obtained weights are retained as data processing tools, and the trained network is deployed as a functional product.

This paper proposes a machine learning method where the Bellman equation is applied in successive waves, across different stages, with varying time differences contributing to continual learning.

Traditional RL solutions for escape-from-room scenarios use only two terms. This paper introduces a three-term formula, adding a step in the future reward search, demonstrating that looking one step further enhances anticipatory intelligence. The approach is based on the policy improvement theorem [32]:

$$\begin{aligned}
 v_{\pi}(s) &\leq q_{\pi}(s, \pi'(s)) \\
 &= E_{\pi'} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_{\pi}(S_{t+2}) \middle| S_t = s \right] \\
 &\leq E_{\pi'} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 S_{t+3} + \gamma^3 v_{\pi}(S_{t+3}) \middle| S_t = s \right]
 \end{aligned}
 \tag{2}$$

4 Methodology

Our proposed methodology uses these three essential elements: i) agent self-sufficiency, ii) the consolidation and secure storage of acquired policies, and iii) ensuring that learned policies are primed to facilitate future learning processes. This section expands on each one of these elements.

4.1 Agent Self-Sufficiency

The neural agent in this study comprises nine sigmoidal neurons that inhibit each other with balanced negative weights and share a common self-activating excitatory input. This shared input is a repetitive ramp K that, in each cycle, forces all neurons to race towards a 1.0 vector output, as illustrated in Fig. 2. Initially, all neuron outputs are set to 0.5. As the ramp grows, energy is consumed until one neuron crosses a preset energy threshold, usually 0.7, at which point it is declared the winner. This neuron determines the board location where the agent places its token. Internal noise ensures the race is balanced and unpredictable, with only one winner.

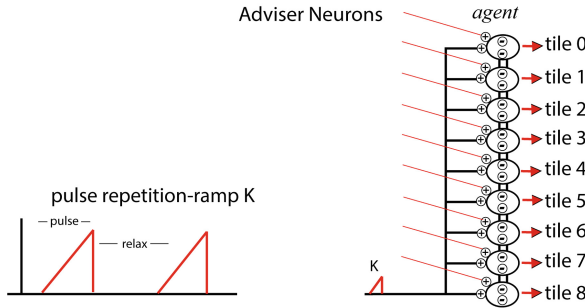


Fig. 2. Neural Agent: This operative block comprises nine sigmoidal neurons that inhibit each other with equal negative weights. A common self-activating excitatory input ramp (K) forces the neurons to race toward 1.0. For each ramp pulse, one random neuron crosses a preset energy barrier (0.7) and is declared the winner, determining the tile for the next move.

Under these balanced racing conditions, connections with external adviser neurons can alter the probability of a specific neuron winning, turning the race into a Markov chain. This energy-consuming network drives the playing robot, creating a dynamic entity that places tokens on the board randomly. To become a skilled player, the agent must learn basic rules and explore game strategies at different stages. The ultimate goal is to memorize rewarding events within the adviser networks' structure using reinforcement learning. The agent's inputs can connect in parallel with the outputs of multiple adviser networks.

4.2 Energy Barrier

Figure 3 shows the actions inside a neuron of the agent. The agent remains inactive when the ramp K is at its lowest value. For stability, all output neurons are initialized at 0.5. At this stage, the adviser's weights and outputs have no immediate effect and remain dormant until K increases and energizes the system. As K grows, its influence gradually manifests. Over time, K increases, and neurons incrementally gain energy. During this period, neurons must cross the threshold, and the connection weights with external neurons will cumulatively affect the race outcome. The governing equation for this process is:

$$P_i = K' \times n \times w_K + w_a(i) \times a_i + \sum_{j=0}^8 out_j \times w_j \quad (3)$$

where,

- K' = the derivative of ramp K , how much the potential increases in each time,
- P_i = internal potential of neuron i ,
- n = time step,
- w_K = connective weight with associated adviser,
- a_i = output of associated adviser neuron,
- out_j = output of any neuron where $j \neq i$,
- w_j = common inhibitory weight.

As time goes by (n grows), the potential inside the neuron P_i forcibly increases until eventually one neuron will cross the energy barrier, set as a hyper-parameter at 0.7. Depending on the size of the derivative of K denoted by K' the average time required to cross the energy barrier is given by the pseudocode:

Thus, for any time n , the neuron i will be influenced by the output of its associated adviser a_i . Probabilities to win the race are influenced by advisers, and the system behaves as a Markov process.

4.3 Adviser Network

Figure 4 depicts a three-layer adviser network incorporating sigmoidal neurons. It receives sparse input code from a webcam and convolutional neural network.

```

n=0;
for  $P_i$  in the agent do
  while  $P_i < 0.7$  do
     $P_i = K \times w_K + a_i \times w_a$ ;
     $K = K + K'$ ;
     $n + +$ ;
  end
end
end
    
```

Algorithm 1: Energy Barrier

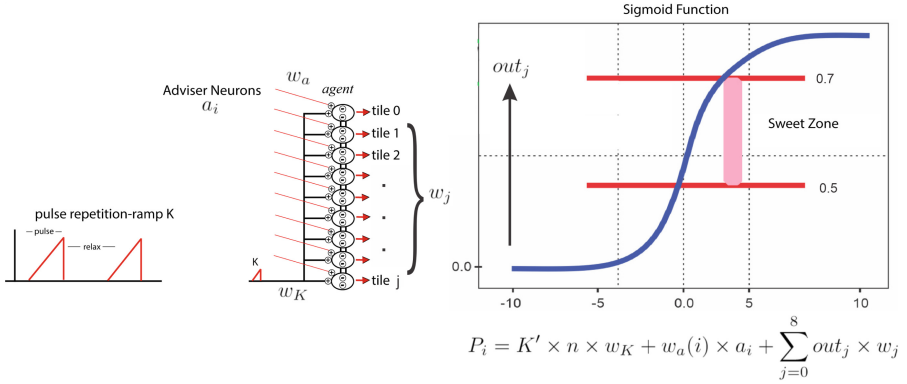


Fig. 3. Energy barrier. The internal potential of all participant neurons is positively modified by a ramp pulse K and a repetitive weight w_k . As time passes by, the potential inside neurons grows, making the neurons race toward 1.0. Eventually, one neuron will cross the energy barrier. During the average time required to cross the sweet zone, the output of the racing neuron depends on $k \times w_k$ and $a \times w_a$. The probability of winning the race depends on the accumulative effects of $out\ a$ for every neuron in the agent.

This network can also utilize labeled data to learn optimal responses for various board states using reinforcement learning and gradient descent. Adviser networks store accumulated policy knowledge essential for developing an advanced playing robot, requiring multiple such networks for comprehensive gameplay modeling.

4.4 Two Stage Continual Learning

Figure 5 shows a two-stage continual learning system. The webcam, in combination with the conv net, feeds sparse data to two adviser neural nets: a_1, a_2 . In the beginning, all weights connecting the output of the adviser with the agent input bus are zero. In this condition, the neural agent can randomly locate its tokens in the board tiles just by burning energy, whether the tiles are preoccupied or not (baby behavior). The process of continuous learning is presented in Fig. 6, where the neural agent generates an adviser a_1 which generates knowledge that is taken by the adviser a_2 , which together with the decision-making agent creates the high-level Markovian Player.

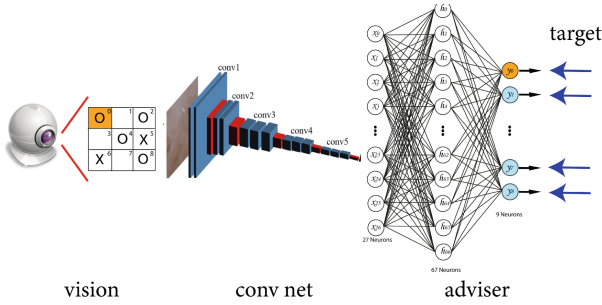


Fig. 4. Typical Adviser Network. A three-layer network with sigmoidal neurons receiving sparse input code. It also receives information about the desired vector target and learns through RL and gradient descent to produce particular responses to particular board states.

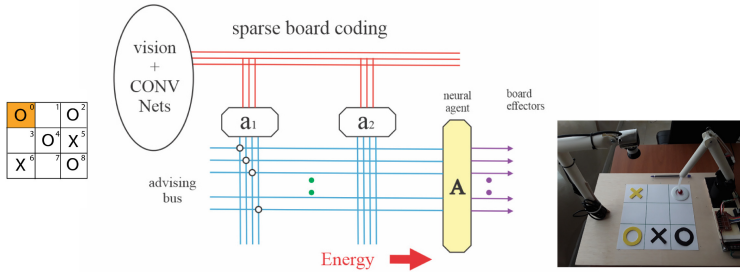


Fig. 5. Two adviser networks aid to master Tic-tac-toe game following continual learning principles. The agent uses adviser a_1 to learn a policy that places tokens only in empty squares. Once this basic knowledge is mastered and safely stored in a_1 , the agent uses a_2 to learn a more advanced game policy strategy.

In the initial stage, the agent trains a_1 to identify occupied squares on the board through exploration and reinforcement learning. Once a_1 masters this task, its weights are frozen, and it becomes a read-only network. Its outputs are then connected to the agent’s input bus with small negative weights, ensuring the agent avoids placing tokens in occupied squares. This sets the stage for stage two, where adviser a_2 learns advanced playing strategies using reinforcement learning and noise-balanced training.

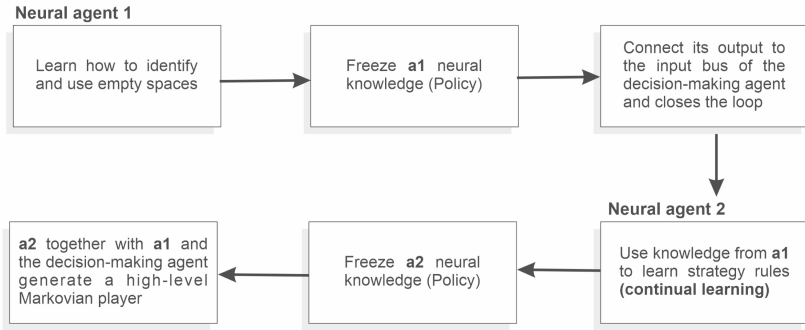


Fig. 6. Flowchart of the continual learning process, in which the neural agent uses knowledge generated by advisor neural networks.

5 Results

5.1 Continual Learning Training

The algorithm sets hyperparameters (5000 training cycles, $w_k = 0.01$, $w_a = 0.05$, neuron gain, learning rate) and undergoes continual learning until a minimal error criterion is met in Stage Two. This process, taking about five hours on a standard desktop, is unsupervised, with the machine learning independently from the initial setup.

5.2 Stage One

Consolidate and keep in a safe place the knowledge that solves the policy of avoiding occupied squares. This stage follows the following steps:

- Learn to recognize occupied squares to stay away from them using the assets of a_1 .
- Run the algorithm until the error decreases below the prefixed threshold.
- Freezes the weights of a_1 and keeps them safe (becomes read-only).
- Initiate learning stage 2.

5.3 Stage Two

The policy learned by a_1 is ready to support the forthcoming learning processes. Thus, Stage Two follows the next steps:

- Use the read-only capacities of a_1 obtained in Stage One.
- Connect the outputs of a_1 to the agent inputs (the matrix diagonal) with small repetitive, negative weights (-0.01).

Now, Stage Two is ready for continual learning. The agent will only place its token in empty tiles, and a_2 is ready to learn game strategies efficiently.

- Use policy learned in Stage One to efficiently explore game situations.
- Use three-term Bellman equation to learn game strategies [9].
- Run algorithm until error decreases under a given threshold.
- Freeze the weights of a_2 that solve the game strategies policy and connect its outputs to agent input (the matrix diagonal) with small repetitive, positive weights (+0.05).

5.4 Experimental Environment

As depicted in Fig. 5, the experimental setup involves an artificial vision system utilizing a webcam connected to a group of convolutional neural networks (conv nets). This setup extracts sparse information about the board state, feeding it into two neural networks, denoted as a_1 and a_2 . Initially, a RL agent employs network a_1 to learn policy P_1 , which ensures the robot places its tokens only on vacant tiles. Once mastered, the weights of a_1 are frozen, and its outputs are routed through an advising bus and weighted connections to facilitate a second RL agent’s learning of policy P_2 using the knowledge embedded in P_1 . Policy P_2 focuses on more intricate game strategies and is stored in network a_2 . After mastering policy P_2 , the weights of a_2 are similarly frozen, and its outputs are connected to the advising bus agent via weighted connections. Through continual learning exploration, the system evolves into a proficient Tic-Tac-Toe player, which subsequently guides an intelligent robot.

The system’s implementation leverages C++ and OpenCV. During a typical operation, the agent initializes with random weights, resulting in random board movements. Through an exploratory phase spanning two stages and numerous self-play games, the agent iteratively refines its strategies until it converges to a minimal error plateau, signaling readiness for gameplay.

Figure 7 illustrates a typical scenario where the system exhibits iterative convergence behavior, achieving consistent success over time. On other hand, Table 1 shows a summary of 500 Tic-Tac-Toe games executed by the trained agent versus an opponent, which presents an excellent performance during the course of the game.

Table 1. Summary of 500 games of the neural agent against an opponent designed to evaluate the performance of the process

	Wins Loses Tie		
Agent	410	20	70
Opponent	20	410	70

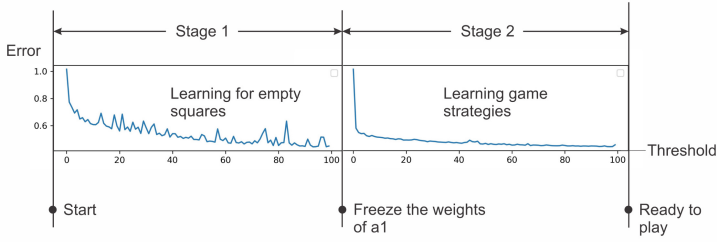


Fig. 7. Convergence behavior on the two stages of the learning process

6 Conclusions

This paper presents a method for achieving effective continual learning in a physical robot equipped with a webcam, internal neural networks, and RL agents. The system autonomously learns to play advanced levels of Tic-Tac-Toe with minimal human intervention, emulating child-like learning strategies. For instance, our system generates the sequence depicted in Fig. 8.

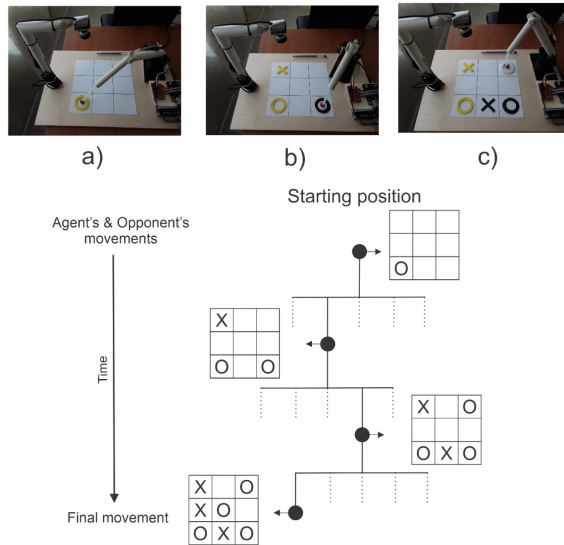


Fig. 8. Sequence of movements using Continual Learning.

By introducing an energy barrier and a configuration of competing neurons, the approach integrates policies learned at different stages, allowing the robot to develop and refine strategies over time. The neural networks securely store

knowledge with frozen weights, enabling the system to act as a valuable advisor in future tasks and demonstrating strategic foresight in gameplay against human opponents. The ultimate contribution involves deploying a novel neural bus channel capable of orderly and extensive interconnection among diverse policy-solving neural networks. This setup fosters a self-converging environment conducive to continual learning, resulting in a robot proficiently performing complex mechanical and logical tasks.

Despite the promising results, this study has certain limitations. The continual learning framework, while effective, requires significant computational resources and fine-tuning of hyperparameters. Additionally, the scope of tasks tested is relatively narrow, and further research is needed to assess the system's performance across a broader range of tasks.

Acknowledgements. This work is part of the R+D+i projects PID2019-108915RB-I00 and PID2022-140907OB-I00 as well as by the grant PRE2020-094056 funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU. It has also been partially supported by Junta de Comunidades de Castilla-La Mancha/ERDF (SBPLY/21/180501/000030) and by the University of Castilla-La Mancha (2022-GRIN-34436).




References

1. Ashfahani, A., Pratama, M.: Autonomous Deep Learning: Continual Learning Approach for Dynamic Environments. In: Proceedings of the 2019 SIAM (2019)
2. Barnett, S., Ceci, S.: When and where do we apply what we learn? A taxonomy for far transfer. *Psychol. Bull.* (2002)
3. Beck, J.: Combinatorial games: tic-tac-toe theory, vol. 114. Cambridge University Press, Cambridge (2008)
4. Bremner, A., et al.: The development of multisensory representations of the body and of the space around the body. *Multisensory Dev.* (2012)
5. Brown, A.L., Kane, M.J.: Preschool children can learn to transfer: learning to learn and learning from example. *Cogn. Psychol.* (1988)
6. Canaan, R., et al.: Leveling the playing field—fairness in AI versus human game benchmarks (2019)
7. Chang, O.: Self-programming robots boosted by neural agents. In: Wang, S., et al. (eds.) *Brain Informatics*. Springer International Publishing (2018)
8. Chang, O., et al.: A protein folding robot driven by a self-taught agent. *Biosystems* (2021)
9. Chang, O., Zhinin-Vera, L.: A wise up visual robot driven by a self-taught neural agent. In: *Future Technologies Conference (FTC)* (2021)
10. Chen, Z., Klahr, D.: All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* (1999)
11. Cichon, J., Gan, W.B.: Branch-specific dendritic ca^{2+} spikes cause persistent synaptic plasticity. *Nature* (2015)
12. DeVries, R., Fernie, D.: Stages in children's play of tic tac toe. *J. Res. Childhood Educ.* (1990)
13. Fong, G.T., et al.: The effects of statistical training on thinking about everyday problems. *Cognitive Psychol.* (1986)

14. French, R.M.: Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* (1999)
15. Golomb, S.W., Hales, A.W.: Hypercube tic-tac-toe. In: *More Games of No Chance, Proceedings of MSRI Workshop on Combinatorial Games* (2000)
16. Kompella, V.R., et al.: Continual curiosity-driven skill acquisition from high-dimensional video inputs for humanoid robots. *Artificial Intell.* (2017)
17. Kosoy, E., et al.: Exploring exploration: comparing children with RL agents in unified environments (2020)
18. Krueger, K.A., Dayan, P.: Flexible shaping: how learning in small steps helps. *Cognition* (2009)
19. Kumaran, D., et al.: What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* (2016)
20. Ling, S.H., Lam, H.K.: Playing tic-tac-toe using genetic neural network with double transfer functions. *JILSA* (2011)
21. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: *Advances in Neural Information Processing Systems*, pp. 6467–6476 (2017)
22. McClelland, J., et al.: Why there are complementary learning systems in the hippocampus and neocortex. *Psychol. Rev.* (1995)
23. McClelland, J.L.: Incorporating rapid neocortical learning of new schema-consistent information. *J. Exp. Psychol. General* (2013)
24. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: the sequential learning problem. In: *Psychology of Learning and Motivation* (1989)
25. Michie, D.: Experiments on the mechanization of game-learning part I. characterization of the model and its parameters. *Comput. J.* (1963)
26. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* (2015)
27. Parisi, G.I., et al.: Continual lifelong learning with neural networks: a review. *Neural Netw.* (2019)
28. Schaul, T., et al.: Prioritized experience replay (2016)
29. Shea, C.: Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychon. Bull. Rev.* (2002)
30. Shin, H., et al.: Continual learning with deep generative replay. In: *Advances in Neural Information Processing Systems*, pp. 2990–2999 (2017)
31. Stein, G., et al.: *The Handbook of Multisensory Processes*. MIT Press (2004)
32. Sutton, R., Barto, A.: *Reinforcement Learning: An Introduction*. MIT Press, Adaptive Computation and Machine Learning series (2015)
33. Zaslavsky, C., et al.: *Tic Tac Toe: And Other Three-in-a Row Games from Ancient Egypt to the Modern Computer*. Crowell, New York (1982)



Centinela: An Intelligent System Based on an Integrated Architecture for Supporting Scholars

Lorena Recalde^(✉), Gabriela Suntaxi, Diana Martinez-Mosquera,
Rommel Masabanda, and Danny Cabrera

Department of Informatics and Computer Sciences, Escuela Politécnica Nacional,
Ladrón de Guevara E11-253, Quito, Ecuador
{lorena.recalde,gabriela.suntaxi,diana.martinez,
rommel.masabanda,danny.cabrera}@epn.edu.ec

Abstract. Research and discovery are fundamental to the advancement of visionary universities. In the modern era, the impact of scientific research on development underscores the critical need for enhanced collaboration among experts. In Ecuador, both public and private universities are spearheading high-quality research initiatives that engage hundreds of scholars, highlighting an increased momentum in scientific output. This surge necessitates an efficient computational tool to support these academic communities in networking, finding collaborators, and selecting joint research topics. To address these needs, we have developed an architecture underpinning a new application named “Centinela”. This architecture is structured into three layers: *i.* two operational and one analytical databases that forms the data layer, which by combining multiple databases, optimizes the retrieval, processing, and presentation of research data; *ii.* Django for the back-end to handle business logic; and *iii.* Angular for the front-end that supports the user interaction. Centinela integrates data from the Scopus API, focusing on the contributions of Ecuadorian researchers, and builds networks of co-authorship, acts as a search engine for academic publications, allows researchers to form groups, provides recommendations on research topics and facilitates group decision-making processes. This integration showcases the application’s robustness in streamlining the management of researchers’ collective expertise and fostering in-depth knowledge across both foundational and advanced subjects.

Keywords: System architecture · Docker · Neo4j · Information retrieval · Group recommender system · Co-authorship network · C4 Model

1 Introduction

In the rapidly evolving landscape of global research and innovation, the integration of advanced technological solutions into academic collaborations is becoming increasingly crucial. These integrations are essential for enhancing the efficiency, accessibility, and effectiveness of research activities. Actually, researchers have social-like online tools that support their work and decisions by either retrieving academic-related information through search platforms or generating recommendations such as academic positions, recently published works relevant to their interests, researchers to follow, and question-answer interactions. These tools foster a vibrant academic ecosystem by bridging the gap between data generation and its practical application, thereby facilitating quicker and more innovative solutions to complex scientific questions. As countries worldwide strive to leverage technology to maximize their research output, the focus shifts towards creating systems that not only accumulate vast amounts of data but also enable meaningful collaborations and insights [9].

Ecuador's scientific output has seen a remarkable expansion over the past decade. As documented in recent studies, the country has contributed over 29,000 publications to international and prestigious journals, showcasing the highest growth in scientific production in Latin America since 2015 [7]. This surge highlights not only the vitality of Ecuador's research landscape but also presents emerging challenges related to research efficiency and collaboration. Despite this growth, Ecuadorian researchers continue to grapple with significant obstacles in accessing timely and relevant information. For instance, the existence of the repository of Ecuadorian academic information, called REDI,¹ remains relatively not visible despite its relevance. On the other hand, traditional approaches—predominantly manual methods such as networking at conferences or utilizing general online social networks—are still prevalent. These methods often result in inefficiencies and a lack of coherence in forming research groups, indicating a pressing need for innovative solutions to foster effective collaboration within the scientific community.

In response to these challenges, this paper proposes a novel integration of two pre-existing platforms designed to promote academic collaboration among Ecuadorian researchers. These platforms, *ResNet* [2] and *ReSearchDecide*,² aim to enhance the efficiency and effectiveness of scientific research in Ecuador. ResNet builds co-authorship networks from scientific publications using data from the Scopus API and is based on an information search and retrieval engine, while ReSearchDecide seeks to provide researchers with functionality to *i*) create research groups, *ii*) receive topic recommendations to start and build their

¹ Ecuadorian Repository of Researchers (REDI) provides data about authors, publications and areas of knowledge, from higher educational institutions. Its main objective is to facilitate the identification of researchers and related areas of knowledge.

² Thesis work authored by Daniel Aimacaña and Renato Padilla (2023), Escuela Politécnica Nacional. More information in <https://biblioteca.epn.edu.ec/cgi-bin/koha/opac-detail.pl?biblionumber=110640> and <https://biblioteca.epn.edu.ec/cgi-bin/koha/opac-detail.pl?biblionumber=110632>.

project proposals and *iii*) have a mediating or consensus system when making group decisions. The integration aims to enhance the operational efficiency and effectiveness of scientific collaboration among Ecuadorian researchers.

A critical barrier to achieving seamless collaboration has been the lack of interoperability between these platforms. The platforms operate in silos, which stifles the exchange of information and impedes collaborative efforts. Addressing this issue, the proposed project, “*Centinela*”, aims to merge these platforms into a unified system that not only facilitates access to relevant information but also significantly enhances the formation and management of research groups. This integration promises to elevate the quality of research and reinforce the international stature of the Ecuadorian scientific community. The project entails a strategic redesign of both platforms to ensure they are interoperable, allowing for efficient communication. The proposed architectural solution features a centralized front-end interfacing with independent services, each residing in its dedicated container, thus enabling robust data exchange through the HTTP protocol. This architecture is pivotal as it not only ensures the transparent flow of information but also supports the dynamic needs of the research community, fostering a responsive and adaptable research environment. This initiative is anticipated to provide substantial benefits not only to individual researchers but also at an institutional level, enhancing strategic decision-making and bolstering the reputation of Ecuadorian research on a global scale. The *Centinela* platform, therefore, is positioned as a critical tool in the advancement of academic collaboration, ensuring that the Ecuadorian scientific community remains at the forefront of global research innovation.

The remainder of the paper is organized as follows: Sect. 2 reviews related literature and contextualizes the present work; Sect. 3 elaborates on the architecture of the system; Sect. 4 outlines i) the preliminary user interfaces which are designed after the *Centinela* software architecture is defined, and ii) the decisions made regarding the prototype development; finally, Sect. 5 presents conclusions drawn from the design and further implementation and outlines potential avenues for future work.

2 Related Work

In this section, we will discuss related work involving academic collaboration platforms, co-authorship networks, and systems designed to improve research group formation and decision-making.

Regarding research group formation, co-authorship networks have been used to study the relationships between authors. In these networks, a node represents an author, and an edge between two nodes indicates that the connecting authors have collaborated on an article. The authors in [1] study the evolution of co-authorship networks. Their work focused on studying the link prediction for creating new connections between authors using machine learning algorithms. In [3], the authors proposed a system to recommend potential collaboration between authors. The recommender system uses the author’s background and

topological features of the network. These works, however, do not provide a visualization and collaboration platform.

Regarding academic collaboration platforms, some proposals are close to Centinela. Barbosa et al. [4] presented an approach for visualizing recommendations based on co-authorship networks called VRRC. VRRC is a web-based tool that allows researchers to identify new possible collaborators. The authors in [6] developed a platform, CollabSeer, which helps discover potential research collaborators based on similarity metrics of the co-authorship network. Other platforms in this line are CNARe [11], VOSViewer [13], and CiteScore [10]. CNARe is an online system that shows personalized collaboration recommendations to researchers. The system allows visualization and analysis of co-authorship social networks. VOSviewer is a widely used software tool designed for constructing and visualizing bibliometric networks. It allows users to create maps based on network data, facilitating the analysis of co-authorship, citation, and co-occurrence relationships within academic literature. CiteScore is a metric used to measure the citation impact of academic journals, providing a broad overview of journal performance by calculating the average citations received per document published in a journal over a four-year period. These platforms, however, do not consider the formation of research groups or the recommendation of research topics, which are core strengths of Centinela.

3 Centinela System Architecture

The proposed architecture for Centinela includes a single centralized front-end that will communicate with two back-ends (independent services), each hosted in its corresponding container and establishing communication with its own database. This design, presented in Fig. 1, ensures interoperability between ResNet and ReSearchDecide, allowing efficient exchange of information among the components facilitated by API links. Then, data flows through the system, allows for real-time data updates and synchronization across the user interface

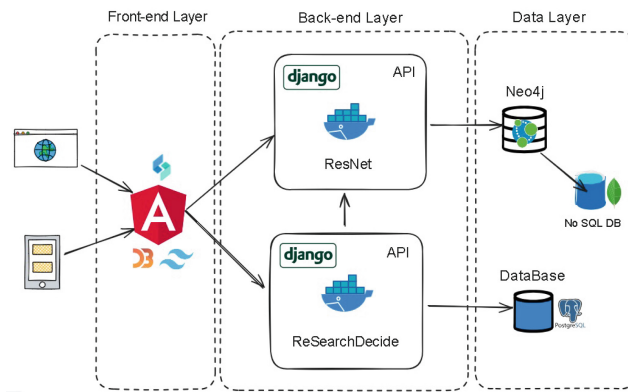


Fig. 1. High Level Infrastructure.

and back-end services. Each component is encapsulated within Docker containers, signifying a modular approach that enhances scalability and system isolation, which is crucial for security and maintenance. A high-level description is presented as follows.

- Front-end Layer. The front-end, represented by Angular, serves as the user interface, accessible via web browsers and mobile devices. This layer is responsible for presenting data in a user-friendly manner and handling user interactions. The front-end is enriched with services deployed in Docker and libraries such as Typescript and Tailwind for layout design as well as D3.js and ngx-charts for creating dynamic, interactive graphics.
- Back-end Layer. Both platforms, ResNet and ReSearchDecide, operate within their respective Docker containers, providing a services architecture. Each service communicates with the front-end via APIs, allowing for robust data processing and business logic execution. ResNet consumes data from a Neo4j graph-oriented database and a non-volatile MongoDB database replica. On the other hand, the ReSearchDecide component interacts with its own dedicated database (PostgreSQL), indicating tailored data handling and storage solutions that optimize its operations for users' accounts.
- Data Layer. The databases connected to ResNet and ReSearchDecide illustrate a distributed data management approach. The use of SQL and Neo4j databases combines both relational and graph database technologies, optimizing the storage and retrieval of structured and relationship-heavy data. While integrating multiple database systems-SQL for structured data handling, Neo4j for complex relationship queries, and MongoDB for large-scale data analytics-presents challenges that include increased complexity in database management, potential performance bottlenecks if not properly optimized, and the need for advanced skills in maintaining data consistency across different types of databases, it also offers extensive benefits. That is, by utilizing SQL databases enables Centinela to efficiently handle structured data, such as user profiles, group memberships and research topic recommendations. This supports complex queries and transactions with high reliability and consistency. By incorporating Neo4j, Centinela excels in constructing and navigating the co-authorship networks, enabling researchers to visualize and explore their academic connections intuitively. MongoDB, allows Centinela to perform data analytics at scale, accommodating vast datasets with flexibility and providing dynamic schemas for rapid application development.

3.1 C4 Model for Software Architecture

The C4 model, conceptualized by Simon Brown, is highly effective in visualizing software architecture through four levels of abstraction [5]. It is particularly suitable for agile environments due to its support for rapid iteration and adaptability. The C4 model includes diagrams and descriptions as presented next.

System Context Diagram (1st Level) This diagram presents how Centinela interacts at a high level with external entities, clearly delineating relationships and dependencies that are crucial for understanding the system’s external interfaces. Figure 2 shows that the external core source of data is Scopus and presents three types of users with their corresponding roles: *Admin users* have administrative privileges and are authorized to execute data updates to Scopus; *Registered users* can view academic information and dashboards and create research groups to receive recommendations and participate in the consensus process; and *Non-registered user* who have limited access to the Centinela system and can only view academic information and dashboards.

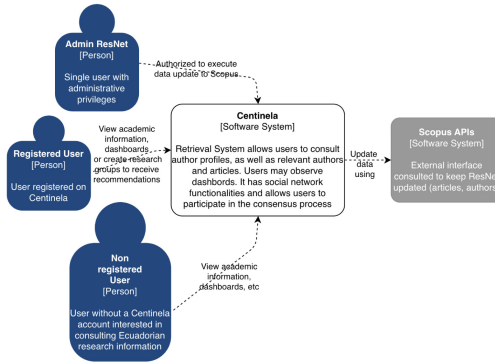


Fig. 2. Centinela’s System Context Diagram.

Container Diagram (2nd Level) The container diagram provides insight into the major functional and technological blocks of Centinela, explaining the communication flows and technological frameworks used in each container. This level of abstraction for the system architecture offers a comprehensive understanding of the system’s structure and modularity. Figure 3 shows the Container Diagram. The two main components are (1) the Web App ResNet, which connects to the Scopus APIs to update information about researchers, and (2) the Web App ResearchDecide, which provides social network functionality and group creation and consensus workshops.

Component Diagram (3rd Level) The Component Diagram (Fig. 4) describes the organization and responsibilities of each component within Centinela’s containers. It details component interactions that facilitate specific functionalities, emphasizing the cohesive and efficient design of the system. The Centinela’s Component Diagram presents the complete integration of ResNet and ReSearchDecide. Next, we provide an overview of the system, highlighting the key components, the interactions between them and how they contribute to strengthening the Centinela platform.

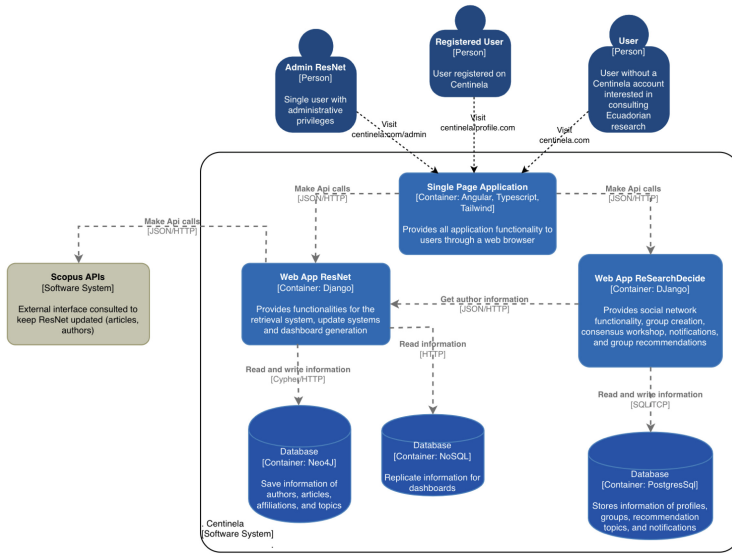


Fig. 3. Centinela’s Container Diagram.

Code Level (4th Level) It describes individual components showing how that component is implemented. A code diagram, e.g., UML class, can be used to zoom into an individual component. The sequence diagrams and description for Centinela, corresponding to Level 4 of the C4 model, provide a granular view of the interactions among specific system components. These diagrams detail the workflow and show how Centinela efficiently handles user interactions during sessions, facilitating a dynamic and responsive collaborative environment supported by robust backend technologies and real-time communications. Figure 5 details the process of user interaction with the Centinela platform during a group consensus session, where members try to decide on a research topic to work on their next project proposal. The process is as follows:

1. Starting the Session: When a user joins Centinela, they connect to a group through a WebSocket managed by the GroupConsumer. This action creates a new session in the database and updates the active user count in real time.
2. Assigning Topics: The user can request research topics, which the system selects and assigns from the database based on the group recommender system. The chosen topics are then sent back to the user.
3. User Searches: Users can search about a topic or a combination of topics on Google search.
4. Adding Topics: Users can add new research topics to the system, but only if no one has moved to the next phase of the consensus process. Once added, the system stores the topic and notifies all group members.
5. Setting Expertise: Users can adjust their expertise level on specific topics, which are saved in the system and shared with the group.

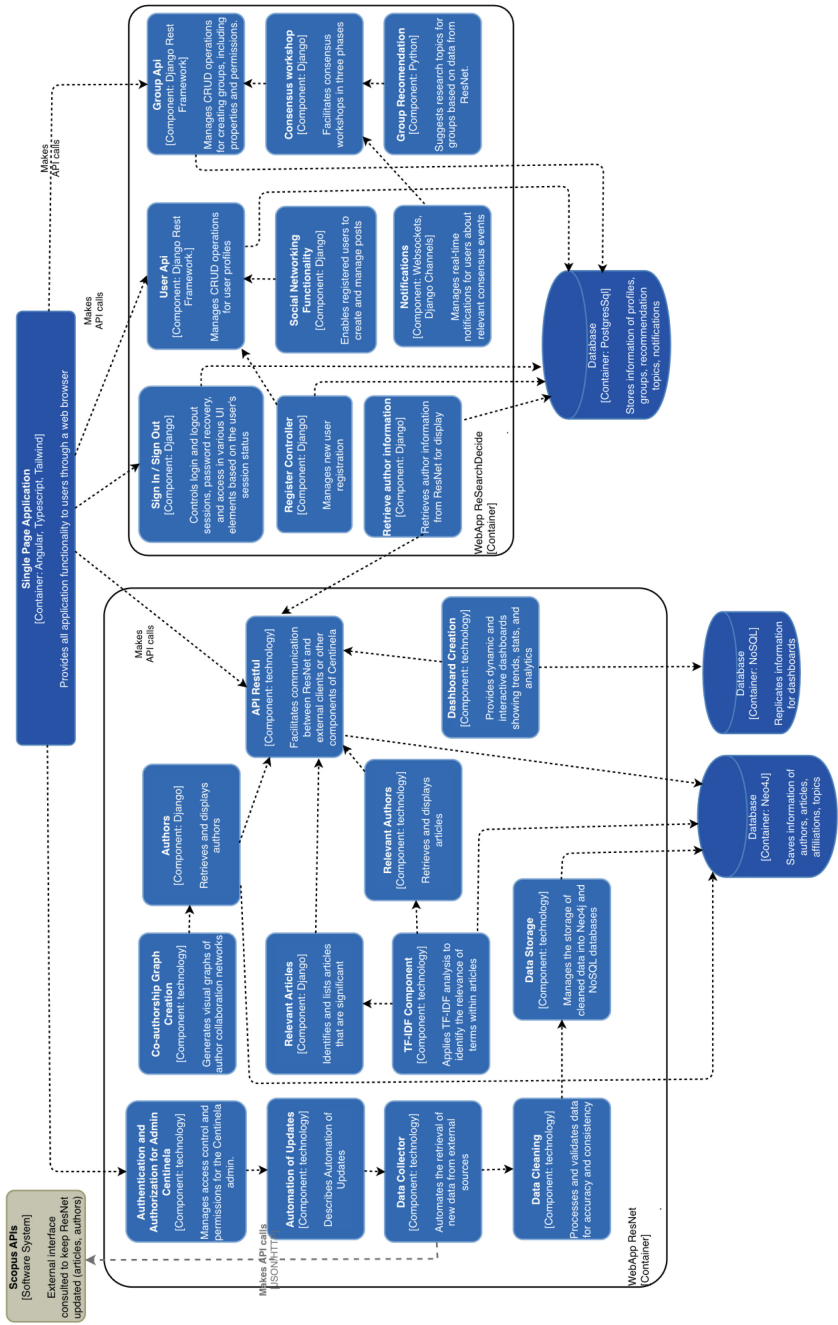


Fig. 4. Centinela's Component Diagram.

6. Moving to the Next Phase: When ready, a user can advance to the next phase. The system updates the phase status and notifies all group members in real time about the progression.

This detailed workflow showcases how the Centinela platform facilitates dynamic and collaborative sessions. There is a WebSocket-based communication model and a series of REST interactions with the server.

3.2 Centinela’s System Components

Next, we describe the main components of the Centinela system.

Information Retrieval Engine (ResNet) This component, ResNet, is responsible for fetching and processing data. It ensures that users have access to up-to-date and relevant scholarly content from Ecuadorian authors, enhancing the utility of the platform for academic research. ResNet applies the TF-IDF technique (frequency-inverse document frequency) in text modeling by working on paper abstracts. TF-IDF plays a fundamental role in Natural Language Processing applications because it is highly effective in improving the accuracy and relevance of the results obtained in the search for relevant authors and articles given a query. The ResNet engine interacts with the *Neo4j Data Source* and *ii. Presentation layer*. The Neo4j Data Source consumes data from the ‘Scopus Data Persistence’ layer and stores cleaned data ingested from Scopus. The Presentation Layer feeds modeled data into the ‘Profile Search,’ ‘Relevant Author Search,’ and ‘Relevant Article Search’ components in the system.

Data Persistence Layers These layers provide a robust and scalable storage solution that supports complex data operations (transactions) and analytics, which are crucial for academic research environments.

- Scopus Data Persistence (operational). This database stores data retrieved from Scopus, a comprehensive database of academic papers. After cleaning, the data is stored in a Neo4j graph-oriented database and employed in the ResNet presentation component. Scopus data aligns and complements a PostgreSQL relational database.
- User Profiles Data Persistence (operational). A more structured relational data model that handles registered users’ profiles and their session logs, as well as groups, topic recommendations, and consensus notifications. This data is stored in a PostgreSQL database.
- Data Lake (Analytical DB). Repository that acts as a centralized data storage that allows the analytics tasks of data at scale. It is managed through a MongoDB database and feeds the dashboards of the system.

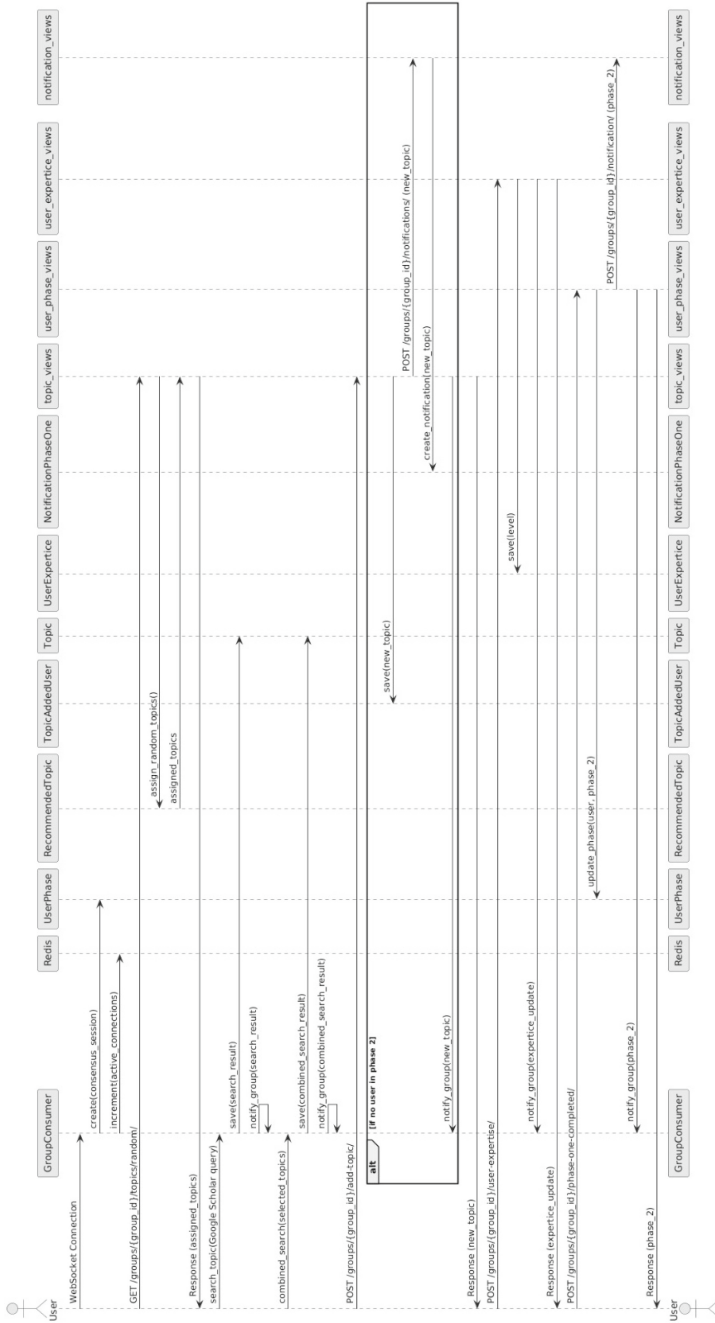


Fig. 5. Centinela’s Code Diagram.

User Interaction Components The user interaction components enhance user engagement and facilitate the discovery of new research opportunities and collaborations. For instance, Profile Search, Relevant Author Search, and Relevant Article Search allow users to search for researcher profiles, as well as authors and articles relevant to a given query. Regarding user interactions, the Data Analytics Dashboards component provides visual analytics to help users understand data trends and derive insights. Likewise, registered users are provided with a social-like platform. That is, the users who create their accounts may handle their scholar profile, create research groups, and generate their posts to the community.

Administrative and Automatic Data Ingest/Update This component automates the collection, ingestion and updating of data to keep the system current and relevant. This implementation reduces manual overhead while trying to obtain Scopus data and ensures the system remains efficient and up-to-date with clean data.

Group Collaboration and Topic Recommender System (ReSearchDecide) On the one hand, this component allows registered users to create research groups and thus promotes a collaborative environment by connecting researchers with common interests. On the other hand, the embedded Group Recommender System suggests potential research topics for the groups based on the members' academic preferences and their past behaviors. Once the topics recommendation is calculated and the list of these potential topics is presented to the group members, the consensus stage takes place. The *Consensus Workshop* includes a consensus model [12] and provides an automated space where the group of researchers can reach a consensus on the list of suggested topics, which supports them in the group decision task of choosing one or two topics from the list for their future joint research proposals.

Actually, among the most important purposes of a Group Recommender System (GRS), there is the consensus or decision-making stage. The decision-making stage also coincides with the frequent activities carried out by a research group, especially when there is a discussion and dialogue to choose an appropriate research topic tailored to the skills and interests of the researchers. Then, grouping researchers and allowing them making decisions require techniques and tools such as the Notification Center. As a significant piece of the consensus stage, the Notification Center keeps group members informed about updates, selections, and other relevant activities within the decision-making actions regarding the topics. The notification center increases system transparency and keeps the research group engaged and informed. Therefore, the consensus workshop functions as a mediator that speeds up the agreement between the members of the group when making a decision and choosing the research topic in which they will be working together. At this point, session data generated by the group members is thought to be useful for further research in the context of GRSs.

User Registration and Management This component ensures secure access to the platform and personalizes user experience based on data-driven insights. It includes *i*) Login/Sign up, which manages user access to the platform; and *ii*) Scholars and Groups Models that stores and manages detailed data about group members (their individual and group profiles), enhancing group functionality and customization.

4 Centinela Prototype

4.1 Centinela GUIs

User feedback has been instrumental in shaping Centinela’s GUIs. Initial reactions highlighted a need for interfaces that could accommodate diverse user preferences and accessibility standards. Feedback on the GUI redesigns was gathered through user testing sessions and iterative feedback loops involving both individual platform components and the Centinela system as a whole. These sessions included 20 participants (registered and potential users) to consider a wide range of perspectives, particularly focusing on usability and functionality alignment with user needs. To quantitatively assess the usability of the redesigned interfaces, the System Usability Scale (SUS) was employed [8]. This standardized scale provided a reliable tool to measure the usability aspects of the Centinela interfaces, guiding our design choices towards enhancing user satisfaction.

The SUS scores obtained helped pinpoint areas needing improvement and confirmed the effectiveness of certain design enhancements. Key areas such as navigation, interactivity, and user engagement received special attention based on SUS feedback, leading to specific adjustments aimed at improving the overall user experience. These enhancements were instrumental in ensuring that the final GUIs not only met the functional requirements of the platforms but also adhered to high usability standards, thereby increasing the acceptance and satisfaction levels among the end users. The adjustments made have considered *i. Customization Features*: based on user feedback, we considered the application usage in diverse kinds of devices and screen sizes to suit their preferences and needs; *ii. Enhanced Navigation*: users reported difficulties in navigating between different functionalities; for example, it was not clear the application navigation for a non-registered user. In response, we redesigned the navigation structure to be more intuitive, with clear labels and logical grouping of features; and *iii. Interactive Elements*: to make the GUIs more engaging, interactive elements such as draggable components, tooltips, and dynamic visualizations were added. These elements make the interface more responsive to user actions and provide immediate feedback on user inputs.

We detail different graphical user interfaces (GUIs) of the Centinela platform, each representing various functionalities and components within the system’s architecture. These GUIs map to specific architectural components and their respective functionalities within Centinela. The design and functionality of these GUIs reflect the system’s emphasis on user-friendly design, robust data management, and interactive, useful visualizations.

Author Search and Network Visualization This interface is a direct manifestation of the Information Retrieval Engine (ResNet) combined with visual displays. Figure 6 shows a co-authorship network, which matches a given query. Users are able to filter and interact with the network based on various parameters like the number of relevant authors and affiliations. This GUI leverages the data persistence layers extensively, particularly using Neo4j for graph-oriented data interactions to map and visualize complex relationships. The system might use advanced querying and graph traversal operations to generate real-time, interactive visualizations that reflect the current state of academic collaborations.



Fig. 6. Searching the relevant co-authorship network for a given query.

User Profile Interface This GUI is closely linked to the User Registration and Management component of the Centinela architecture (Fig. 7). It displays detailed profiles of users, including personal information, institutional affiliations, website links, Scopus ID, citation details, and a list of academic articles indexed in Scopus. It also allows users to edit their profiles. This interface interacts with the back-end layers, where user data is managed and stored in the PostgreSQL and Neo4j databases, depending on the user’s information. It requires robust data handling to ensure that user information is up-to-date and securely managed. This section of the system likely leverages user authentication and data retrieval services to populate and update user profiles.

Social Interaction Interface This GUI pertains to the User Interaction Components. As shown in Fig. 8, it facilitates social interactions similar to academic social networking, where users can post updates, share information about their research projects and multimedia content, and engage with the community. This interface is powered by components that handle dynamic content generation and

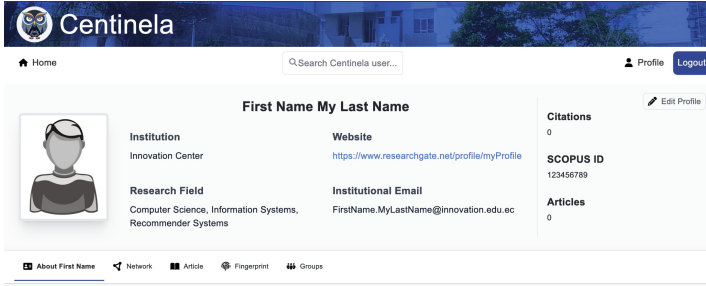


Fig. 7. User profile after registration.

RECALL	text_clean			text_lemma			text_stem		
	ROS	RUS	SMOTE	ROS	RUS	SMOTE	ROS	RUS	SMOTE
SVM	96.6%	85.5%	96.7%	92.1%	85.5%	92.6%	89.7%	77.7%	89.7%
RandomForest	95.7%	77.3%	94.4%	93.6%	72.4%	91.4%	92.6%	63.0%	90.6%
XG-Boost	85.0%	77.6%	85.2%	79.6%	70.5%	79.9%	72.3%	57.4%	73.7%
Naive Bayes	76.0%	73.7%	76.5%	68.8%	66.0%	69.8%	68.8%	61.5%	62.6%
ANN	96.9%	85.1%	96.2%	93.3%	73.0%	93.2%	80.8%	60.4%	92.1%

Fig. 8. A post in the timeline of a registered user.

social interactions, which are managed by the PostgreSQL database. This component ensures that user-generated content is properly indexed and searchable within the network.

4.2 Considerations for Implementing Centinela Prototype

The integration of Centinela's two core systems, ResNet and ReSearchDecide, posed significant technical challenges, primarily around data interoperability and system communication. The former system focuses on information retrieval from academic publications, while the latter facilitates group decision-making and research topic recommendations. The challenge was ensuring efficient data exchange and functionality between these distinct but interconnected systems. In the implementation process of the system prototype, we have considered the next strategies that have not only solved the primary integration challenges but also enhanced the overall robustness of the Centinela functional prototype.

- API Development. We designed robust APIs to facilitate effective communication between the systems. These APIs handle data requests and responses in a standardized format, ensuring that each system can interpret and process the data correctly.

- Data Management. Centinela integrates multiple database technologies to manage a diverse array of data types and structures. Centinela employs Neo4j, PostgreSQL, and MongoDB to support its advanced features, including network visualizations, user management, and dynamic content handling. We selected these databases by considering the challenges of integrating heterogeneous data systems and the solutions needed to enhance data integrity, security, and performance. The required data models were implemented to overcome interoperability issues and reduce data loss during interactions.
- Continuous Integration (CI) Pipelines. To manage integration challenges, CI pipelines were established. These pipelines allow for frequent updates to be made across systems with minimal disruption, ensuring that integration points are consistently maintained and improved upon.
- Modular Design. The containerized approach of Centinela’s architecture allows for modules to be added or updated without affecting the entire system. This design is crucial for adapting to new research requirements or integrating additional functionalities.

4.3 Discussion

Centinela represents a significant advancement in the field of academic collaboration systems. It is specifically tailored to meet the dynamic needs of researchers in Ecuador. Centinela’s architecture has innovative aspects, as described next.

- Novel Architecture Design. Centinela’s architecture is a pioneering integration of information retrieval and group recommender systems, specifically designed to overcome the traditional barriers to academic collaboration. This architecture not only supports the functional requirements of the system but also ensures that all interactions between these components are streamlined and effective, enabling real-time updates and synchronization across the system’s front-end and back-end services.
- GUIs Design and Implementation. Centinela’s GUIs were crafted to enhance user experience, making them both intuitive and engaging for researchers. The design process focused on user-centered principles, ensuring that the interfaces are not only visually appealing but also functional. These GUIs serve as a critical bridge between the complex back-end processes and the users, providing them with easy access to sophisticated research tools and data. The implementation of these interfaces, which were evaluated, demonstrates Centinela’s commitment to enhancing the productivity of its users.
- Sociotechnical Considerations. A deep understanding of the sociotechnical dynamics within academic communities guided Centinela’s development. The platform is designed to enhance user interaction and foster a collaborative environment conducive to shared research efforts. By addressing both the social and technical aspects of collaboration, Centinela helps build more cohesive research groups and improves the overall quality of research output. It has the ability to support diverse researcher interactions and activities.

- Scalability and Adaptability. Centinela’s architecture was conceived with scalability and adaptability at its core. It uses Docker containers to ensure that each component can be independently scaled and updated without disrupting the entire system. Then, Centinela adapts to varying demands and can be easily expanded to include new functionalities. The flexible architecture ensures that Centinela can grow alongside the evolving needs of the academic community, making it a long-term solution for researchers.

5 Conclusion

An increase in scientific production has been observed, which has generated the need for an efficient computing tool to support Ecuadorian communities of scholars in activities like forming networks, searching for collaborators, and analyzing and choosing research topics. The primary objective of this project is to address this problem by integrating the ResNet and ReSearchDecide systems into a single platform called “*Centinela*”. The project focuses on redesigning the two existing platforms to ensure they are interoperable, allowing effective communication between them. The integration of both systems into the unified Centinela platform significantly facilitates access to relevant information and enhances research collaboration capabilities. By merging these systems, Centinela not only streamlines the flow of information with search and retrieval modules, but also creates a more cohesive environment for researchers to engage and collaborate by forming research groups and supporting them in decision-making processes. In this way, the integration of these components into a cohesive platform significantly enhances the capability of researchers to conduct high-quality, collaborative research. With the future launch of Centinela, we hope that it will contribute to the advancement of academic knowledge and innovation, which will strengthen the reputation of the Ecuadorian scientific community both nationally and internationally.

In technical terms, the Centinela architecture is designed to support scientific information access, data persistence, user interaction, and group collaboration. Centinela leverages Scopus’s up-to-date data collection, providing tools for effective data analysis and collaboration. The system’s modular design ensures flexibility and scalability, accommodating evolving research needs and technologies. Indeed, the proposed architecture not only supports the operational needs of Centinela but also aligns with best practices for modern application development, including the use of containers for deployment, services for system decomposition, and APIs for service integration. This setup is ideally suited for an academic collaboration platform like Centinela, which requires robustness, flexibility, and scalability to handle the dynamic and complex data associated with academic research and collaboration. The system meets the current needs of Ecuadorian researchers and is also prepared to adapt to future challenges and opportunities in the academic domain.

As we continue to develop and refine Centinela, our next steps are related to system deployment for public use. Our aim is to offer Centinela as a free

access tool, particularly benefiting public universities and the broader academic community in Ecuador. This initiative is designed to democratize access to advanced research collaboration tools, contrasting with proprietary systems such as the [Research Information Management System](#) offered by Elsevier. We are in the process of establishing the necessary infrastructure to ensure that Centinela is not only accessible but also robust and secure for all users. Deployment plan includes applying key performance metrics such as response time, system uptime, and transaction throughput, which must be continuously monitored. Further details regarding the URL and access protocols will be announced as we approach the official release, ensuring that the academic community can fully leverage Centinela's capabilities for research collaboration and management.

Acknowledgements. The authors gratefully acknowledge the financial support provided by the Escuela Politécnica Nacional for TICEC2024.




References

1. Affonso, F.M., de Oliveira Santiago, M., Dias, T.M.R.: Analysis of the evolution of scientific collaboration networks for the prediction of new co-authorships. *Transinformação* (2022). <https://doi.org/10.1590/2318-0889202234e200033>
2. Arias, J., Recalde, L.: Search and visualization of researcher networks: co-authorship in Ecuador. In: *Conference on Information and Communication Technologies of Ecuador*, pp. 448–463. Springer (2023)
3. Aslan, S., Kaya, M.: A hybrid recommendation system in co-authorship networks. In: *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5 (2019). <https://doi.org/10.1109/IDAP.2019.8875989>
4. Barbosa, E.M., Moro, M.M., Lopes, G.R., Oliveira, J.: Vrrc: web based tool for visualization and recommendation on co-authorship network (abstract only). In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (2012). <https://doi.org/10.1145/2213836.2213975>
5. Brown, S.: *Software architecture for developers. Coding the Architecture* (2013)
6. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: Collabseer: A Search Engine for Collaboration Discovery, pp. 231–240 (2011). <https://doi.org/10.1145/1998076.1998121>
7. Herrera-Franco, G., Montalván-Burbano, N., Mora-Frank, C., Bravo-Montero, L.: Scientific research in Ecuador: a bibliometric analysis. *Publications* **9**(4), 55 (2021)
8. Kaya, A., Ozturk, R., Altin Gumussoy, C.: Usability measurement of mobile applications with system usability scale (SUS). In: *Industrial Engineering in the Big Data Era: Selected Papers from the Global Joint Conference on Industrial Engineering and Its Application Areas, GJCIE 2018, June 21–22, 2018, Nevsehir, Turkey*. pp. 389–400. Springer (2019)
9. Lee, S., Bozeman, B.: The impact of research collaboration on scientific productivity. *Soc. Stud. Sci.* **35**(5), 673–702 (2005)
10. Teixeira da Silva, J.A., Memon, A.R.: Citescore: A cite for sore eyes, or a valuable, transparent metric? *Scientometrics* **111**(1), 553–556 (2017)
11. de Sousa, G.A., Diniz, M.A., Brandão, M.A., Moro, M.M.: Cnare: Co-authorship networks analysis and recommendations. In: *Proceedings of the 9th ACM Conference on Recommender Systems* (2015). <https://doi.org/10.1145/2792838.2796553>

12. Tran, T.N.T., Felfernig, A., Le, V.M.: An overview of consensus models for group decision-making and group recommender systems. In: *User Modeling and User-Adapted Interaction*, pp. 1–59 (2023)
13. Van Eck, N.J., Waltman, L.: Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics* **111**, 1053–1070 (2017)



Optimizing Predictive Models in Healthcare Using Artificial Intelligence: A Comprehensive Approach with a COVID-19 Case Study

Juan Pablo Astudillo León^{1,2}(✉) , Kevin Chamorro¹ ,
and Santiago J. Ballaz³ 

¹ School of Mathematical and Computational Sciences, Yachay Tech University, Urcuquí 100115, Ecuador

² Communication Networks and Intelligent Services Research Group (ComNet Innova YT), Yachay Tech University, Urcuquí 100115, Ecuador
jastudillo@yachaytech.edu.ec

³ School of Biological Sciences and Engineering, Yachay Tech University, Urcuquí 100115, Ecuador

Abstract. The ongoing global healthcare challenges underscore the need for accurate and efficient predictive tools capable of assessing disease severity. This study introduces a broadly applicable artificial intelligence (AI) framework designed to enhance disease severity predictions, focusing on analyzing feature importance and optimizing hyperparameters using the Hyperband method. Utilizing a dataset of 1215 patients, we employed L1 regularization to pinpoint a minimal yet highly informative set of biomarkers—oxygen saturation (O2SAT), partial pressure of carbon dioxide (PCO2), age, percentage of lymphocytes, and percentage of neutrophils. These biomarkers were found sufficient to predict disease severity with 95% accuracy. The Hyperband method facilitated efficient and effective tuning of our neural network models, enhancing their predictive capabilities. Although initially developed for COVID-19, our streamlined model can be adapted to other diseases with similar data characteristics, thereby improving predictive efficiency and impacting healthcare resource allocation and patient management during pandemics. Our findings demonstrate the potential of targeted AI applications to not only refine response strategies during health crises but also provide insights that could lead to more informed and effective healthcare practices globally.

Keywords: COVID-19 · Prognosis · Artificial intelligence · Feature importance analysis · Hyperparameter optimization

1 Introduction

The ongoing challenges posed by pandemics, exemplified by the coronavirus disease pandemic that first emerged in November 2019 (COVID-19), which has claimed over 7,004,638 lives as of March 10th, 2024 [1], underscore the need for advanced predictive tools in healthcare. COVID-19, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has seen evolving viral variants that complicate pandemic countermeasures due to antibody neutralization escape and enhanced transmission [3]. In this environment, early and accurate prediction of disease severity is crucial for effective patient management and resource allocation [14]. Artificial Intelligence (AI) has emerged as a potent tool in managing such pandemics [38], offering data-driven insights that augment patient triage [6, 18]. Recent advancements in AI and machine learning (ML) technologies demonstrate significant potential in leveraging clinical and laboratory data for predictive modeling. These tools are increasingly capable of anticipating disease progression and outcomes with remarkable accuracy across various conditions [16], not limited to COVID-19. Identifying key clinical features that predict disease severity is a major challenge, given the vast array of potential biomarkers and patient data [4]. Moreover, integrating AI into healthcare workflows requires not only accurate but also interpretable and feasible models [5, 20].

This study contributes to the field of AI in healthcare by focusing on optimizing disease severity prediction models through feature importance analysis [40] and hyperparameter optimization [7]. By employing L1 regularization [39], we identified key biomarkers that significantly impact predictions, streamlining the model to enhance both its performance and practical applicability. Additionally, the Hyperband method [25] for hyperparameter optimization was utilized to refine our neural network models, thereby increasing their predictive accuracy while maintaining computational efficiency. Although initially developed for COVID-19, the methodologies and insights derived from this study have broader applications, potentially aiding in the management of other diseases with similar dataset characteristics. The goal is to refine predictive modeling techniques, focusing on the identification of critical biomarkers and the optimization of model parameters, to enhance predictive accuracy and operational efficiency in clinical settings.

2 Methodology

Our approach is designed to refine predictive modeling techniques, initially targeting COVID-19 severity but applicable to any disease where the dataset shares similar characteristics. We focus on identifying critical biomarkers and optimizing model parameters to enhance both predictive accuracy and operational efficiency in clinical settings. Figure 1 illustrates the methodology used in our study, systematically structured into several phases: dataset compilation and preprocessing, feature selection via L1 regularization, model development leveraging neural networks, and rigorous hyperparameter optimization.

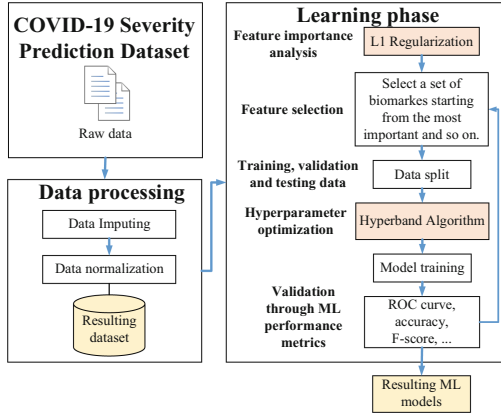


Fig. 1. Methodology employed in our study

2.1 Ethics

Written informed consent was waived, since this study carried out nonexperimental research with secondary data, anonymized patients, type and design with descriptive characteristics. We received ethical approval from the board of the IESS COVID-19 Hospital Quito Sur Ethics Committee. The study was conducted in compliance with national ethical standards and the 2013 revision of the Declaration of Helsinki.

2.2 Dataset Description

This study leverages a comprehensive dataset initially including 4009 patients diagnosed with COVID-19 via molecular tests (CDC 2019–Novel Coronavirus RealTime RT-PCR Diagnostic Panel in upper and lower respiratory specimens) and COVID19-like symptoms upon admission, spanning from March 13 to June 17, 2020. This period marked the beginning of the pandemic in Quito, Ecuador. All the basic information and laboratory results were downloaded from patients’ electronic medical records (IBM AS-400).

We conducted a retrospective analysis focusing on hematological and arterial blood gas parameters. Drawing samples for routine laboratory tests took place upon admission, whereas arterial blood gas exam (gasometry) was performed after a 40 min interval of average, using the Sysmex XN-550 TM (Sysmex America Inc., USA) and the RAPIDPoint B500 (Siemens Healthcare GmbH; Germany) systems, respectively. For the analysis, we systematically categorized patients into “Severe” and “Non-Severe” groups based on their Pneumonia Severity Index (PSI) [12] and partial oxygen pressure [PaO_2] levels [23]. Only those patients with a PSI above 3 (1215) in the triage evaluation were submitted to the gasometry. From them, 489 patients suffered from blood hypoxemia

(PaO₂ < 60 mmHg) and were classified as “Severe” [23]. The rest, 726 patients with a PSI > 3 and PaO₂ ≥ 60 mmHg were considered “Non-Severe”. This stratification facilitated a more nuanced understanding of the disease’s impacts and supports the refinement of predictive models for patient outcomes.

Data Preparation Table 1 shows the COVID-19 severity prediction dataset, comprising data from 1215 patients and 64 key indicators relevant to COVID-19 severity. These indicators, encompassing a wide array of clinical and laboratory variables, provide a robust basis for the development of predictive models. To deal with missing data in our epidemiological studies, we used the `SimpleImputer` from the `scikit-learn` library. We chose the `most_frequent` strategy to replace missing values with the most common value (mode) for each variable. This method is backed by research and helps maintain the consistency of the dataset. By minimizing potential biases introduced by missing data, we can create a more complete dataset for training our machine learning model, improving its accuracy in making predictions.

Table 1. Sample of the COVID-19 Severity Prediction Dataset

O2SAT	PCO2	Age	Lymphocytes pct	Neutrophils pct	...	Class_Severity
84.0	37.0	28	15.3	77.4	...	1
92.0	37.0	69	30.0	62.3	...	0
87.0	23.0	72	16.6	78.4	...	1
94.0		70	8.5	81.5	...	0
91.0	26.0	52	14.7	77.6	...	1

2.3 Feature Processing and Extraction

In our analysis, we emphasize processing and extracting features. We use L1 regularization [27], also known as Lasso regression, to identify the most important markers that predict the severity of COVID-19. L1 regularization aims to reduce the sum of the absolute differences between the actual and predicted values, helping us to find the most relevant and predictive features. This is mathematically represented as follows:

$$\frac{1}{n} \sum_{k=1}^N (y_i - \hat{y}_i)^2 \cdot \alpha \cdot \frac{1}{n} + \sum_{k=1}^N |w_i| \tag{1}$$

where y is the actual output variable and \hat{y} is the expected output, w is the weight, and α is the regularization parameter that it is going to be adjusted.

We created a visual representation using a bar graph, ranking features from most to least important (Fig. 2). This visualization revealed key findings that have implications for both research and clinical decision-making in this context. Firstly, the prominence of oxygen saturation (O2SAT) and partial pressure

of carbon dioxide (PCO₂) as the most critical features underscores the importance of arterial blood gasometry indicators in assessing COVID-19 severity. This emphasizes the crucial role of respiratory function and gas exchange in patients affected by the SARS-CoV-2 virus, consistent with the COVID-19 nature as a viral respiratory disease.

Age emerges as a significant predictor, aligning with clinical and epidemiological evidence suggesting that older patients are at increased risk of developing severe COVID-19. This reinforces the necessity of considering age as a critical factor in risk assessment and clinical decision-making. Additionally, the inclusion of specific laboratory markers such as C-reactive protein (CRP), hematocrit (HCT), glucose, and bicarbonate anion (HCO₃⁻) among the top predictors highlights their relevance as biomarkers of COVID-19 poor progression and their potential utility in predicting severe COVID-19. These biomarkers may reflect the patient’s systemic inflammatory response, electrolyte balance, and metabolic status, which are critical in COVID-19 pathophysiology. Interestingly, the finding that some biomarkers, such as basophil total counts and percentage, as well as the medium platelet volume-to-platelet count (MPV/PC) ratio, hold zero importance suggests that these parameters may not help predict COVID-19 severity within the context of this study. This can aid in simplifying predictive models by excluding variables that do not significantly contribute to model accuracy.

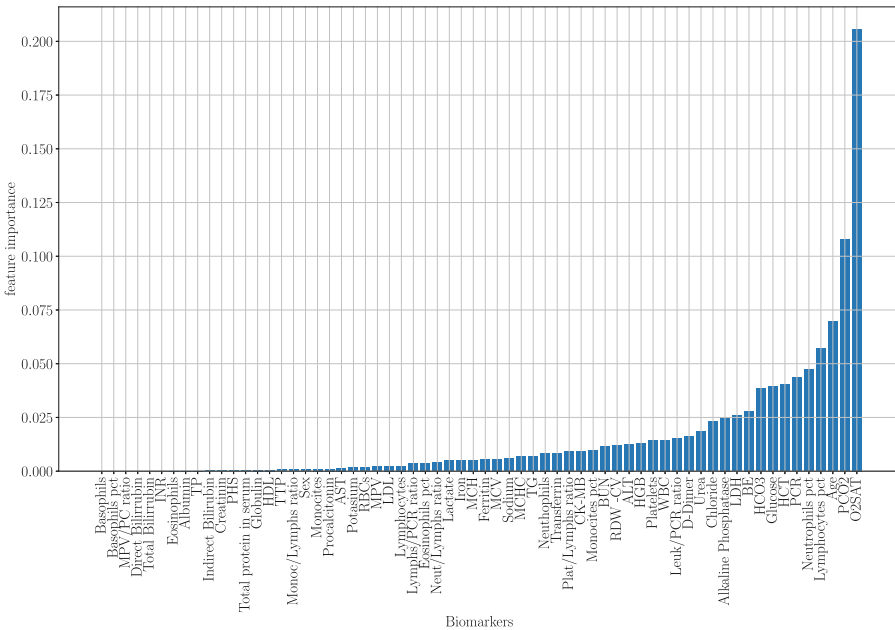


Fig. 2. Distribution of Feature Importance for COVID-19 Severity Prediction

The results suggest that more efficient and cost-effective predictive models could be developed by focusing on a subset of highly predictive indicators. This

could enhance the accuracy of predictions and optimize resource use in clinical settings, especially under high-demand or limited resource scenarios. These highlighted points underscore the complexity of COVID-19 and the relevance of a multidimensional approach to its assessment and management. Identifying the most predictive features provides valuable directions for future research and the development of clinical tools to predict and manage severe COVID-19 disease more efficiently.

2.4 Learning Phase

The Learning Phase of our study was centered on the application of supervised learning algorithms to predict COVID-19 severity based on the refined set of features obtained from the Feature Processing and Extraction phase. Our approach involved training, validating, and testing models to ensure that predictions were robust and reliable.

Model Selection In the process of selecting an appropriate model for our supervised learning task, we concentrated on the characteristics of our dataset, which consisted of numerical values and the binary classification of COVID-19 severity into “Severe” and “Non-Severe” cases. Given these considerations, we decided to utilize neural networks as our primary modeling technique. Neural networks are a class of machine learning algorithms inspired by the structure and function of the human brain. At their core, neural networks consist of layers of interconnected nodes or “neurons,” where each connection can transmit a signal from one node to another. The strength of these connections, or “weights,” is adjusted during the training process to minimize the difference between the predicted outcomes and the actual outcomes. This allows the network to learn complex patterns in the data. The utility of neural networks in our study lied in their remarkable ability to model non-linear relationships and interact with a vast amount of variables, making them highly effective for classification tasks.

2.5 Hyperparameter Optimization

In our study, hyperparameter optimization is critical step in fine-tuning our neural network model to achieve optimal performance. We have chosen to utilize the Hyperband algorithm [25] for this purpose. Hyperband is a modern optimization technique that makes finding the ideal settings for any model more efficient than traditional approaches like Grid Search or Random Search. It surpasses Grid Search, which tries every combination of hyperparameters, and Random Search, which randomly selects combinations without attention to past outcomes. Hyperband is innovative because it changes the number of configurations it explores based on their performance, using resources wisely.

The Hyperband Algorithm Hyperband is a powerful tool for optimizing hyperparameter tuning. It distributes resources wisely to configurations that

have potential and eliminates low-performing models quickly. This “successive halving” method ensures that computational resources are focused on the most promising models, which is crucial for training complex deep learning models with lengthy training times.

Hyperband’s core principle (Algorithm 1) involves a series of tournaments between different configurations. It initiates with a wide range of random configurations, which are then progressively refined. In each round, promising configurations receive additional resources. This iterative approach enables Hyperband to effectively navigate the hyperparameter space and pinpoint the most impactful configurations, all while minimizing computational costs compared to exhaustive methods like grid or random search.

Algorithm 1 Hyperband Algorithm

```

1: Input: max_epochs, factor
2: Initialize:  $s\_max = 1 + \log_{\text{factor}}(\text{max\_epochs})$ 
3: for  $s \in \text{reversed}(\text{range}(s\_max + 1))$  do
4:    $n = \text{int}(\text{ceil}(1 + \log_{\text{factor}}(\text{max\_epochs}) / \log_{\text{factor}}(\text{factor})))$ 
5:    $r = \text{max\_epochs} * \text{factor}^{-s}$ 
6:   Sample  $n$  configurations
7:   for  $i \in \text{range}(s + 1)$  do
8:     Train each configuration for  $r * \text{factor}^i$  epochs
9:     Keep the top  $n/\text{factor}$  configurations

```

Hyperband determines the highest number of iterations (s_max) based on the total epochs and a reduction factor. Each iteration, known as a “bracket,” begins by randomly selecting a set of configurations and training them for a certain number of epochs. Configurations are then evaluated, and only the top performers are selected for further training in the next round. This halving continues until the most promising configuration is identified. The factor controls the rate of reduction in the number of configurations and the increase in the number of epochs for each round, balancing the exploration of the hyperparameter space with the depth of training for promising models.

Neural Network Architecture for Hyperband Optimization To implement Hyperband, we designed a base neural network architecture with adjustable hyperparameters. Table 2 shows the architecture’s flexibility, which allows Hyperband to explore a wide range of configurations to identify the optimal setup. The choice of the sigmoid activation function for the output layer was due to its effectiveness in binary classification problems, producing outputs in the range $(0, 1)$, ideal for modeling the probability of the “Severe” or “Non-Severe” classes. Adam optimizer [17] has been proven to be a good choice compared to the classical stochastic gradient descent procedures. The binary_crossentropy loss function was used for its suitability in binary classification tasks, directly optimizing for the model’s ability to predict the probability of class membership accurately.

Table 2. Base Neural Network Architecture Parameters

Parameter	Range
Input Layer Neurons	8 to 128 (step of 8)
Hidden Layer Neurons	8 to 128 (step of 8)
Number of Hidden Layers	1 to 5 (step of 1)
Dropout Rate	0 to 0.5 (step of 0.1)
Activation Functions	ReLU, Leaky ReLU
Output Activation Function	Sigmoid
Learning Rate	1e-2, 1e-3, 1e-4
Batch Size	2, 16, 32, 64, 128
Optimizer	Adam
Loss Function	binary_crossentropy

Furthermore, an important consideration in neural network design is the prevention of “dead neurons,” which can occur with the use of certain activation functions like ReLU (Rectified Linear Unit). ReLU, commonly used in neural networks, can cause “dead” neurons where the output is always zero, limiting the model’s learning capability. This happens when ReLU’s output is zero for negative inputs and the neuron’s weights adjust to keep the output negative. Instead of ReLU, we explored also Leaky ReLU, which adds a small positive gradient for negative inputs. Unlike ReLU’s zero output for negative values, Leaky ReLU produces a small non-zero output, addressing the issue of dead neurons and enhancing the model’s ability to learn data patterns.

To prevent overfitting, we integrated also Dropout into our base neural network architecture. Overfitting happens when a model captures too many details in the training data, leading to poor performance on new data. Dropout addresses this by randomly deactivating some neurons during training. This forces the model to learn features that are not dependent on specific neurons, creating a more robust model with better generalization abilities.

Objective and Configuration of Hyperband Our study used validation accuracy as the performance measure for Hyperband, seeking to enhance the model’s accuracy on the validation dataset. To ensure a comprehensive search while considering available computational resources, we set *max_epochs* to 10 and *factor* to 3. The number of iterations for Hyperband is determined by a specific mathematical formula:

$$1 + \lceil \log_{\text{factor}}(\text{max_epochs}) \rceil \tag{2}$$

By using Early Stopping, we improved our Hyperband approach. Early Stopping stops training on models that have not improved their accuracy *val_accuracy* after a set number of epochs. This not only saves computational costs but also prevents overfitting.

2.6 Machine Learning Performance Metrics

To evaluate our neural network’s ability to predict COVID-19 severity, we applied several established machine learning metrics:

- **Accuracy:** Reflects the overall model correctness:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3)$$

where TP , TN , FP , and FN stand for true positives, true negatives, false positives, and false negatives, respectively.

- **Precision:** Measures the accuracy of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (4)$$

- **F1-Score:** The harmonic mean of precision and recall, suitable for uneven class distributions:

$$\text{F1-Score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (5)$$

- **AUC (Area Under the ROC Curve):** Indicates overall performance across all classification thresholds, where 1 represents perfect accuracy.
- **Cohen’s Kappa:** Assesses the agreement level between predictions and actual classifications:

$$\text{Cohen’s Kappa} = \frac{p_o - p_e}{1 - p_e} \quad (6)$$

where p_o is observed agreement and p_e is expected agreement under independence.

3 Results and Discussion

3.1 Training Phase

The training phase of our study was guided by the insights gained from L1 regularization, which highlighted the significance of certain metrics in predicting COVID-19 severity [2]. Leveraging the Hyperband algorithm, we embarked on a sequential search to identify the optimal neural network architecture, starting with the most critical metric, oxygen saturation (O2SAT). This approach allowed us to systematically explore the utility of adding more metrics, such as the combination of oxygen saturation (O2SAT) and partial pressure of carbon dioxide (PCO2), and so on, incrementally incorporating additional metrics to evaluate their collective impact on model performance. The aim was to ascertain the minimum set of metrics necessary for accurately diagnosing COVID-19 severity, thereby providing a streamlined diagnostic tool for medical practitioners that focuses only on the most informative indicators [2, 15, 21, 32, 33].

Our dataset was divided into training, validation, and testing sets, comprising 80%, 10%, and 10% of the data, respectively. This division is crucial for several reasons. Primarily, it ensures that the model is trained on a large portion of the data (training set), while also being fine-tuned and validated against unseen data (validation set) to prevent overfitting. The testing set, also unseen during training, is used to evaluate the final model's performance, offering an unbiased assessment of its predictive capabilities. Such a split is essential for developing a model that generalizes well to new, unseen data.

Furthermore, all data were normalized to fall within a range of 0 to 1. Normalization is a critical preprocessing step in machine learning, particularly in neural network models, as it ensures that all input features contribute equally to the model's prediction [35]. Without normalization, features with inherently larger scales dominate the learning process, potentially biasing the model and impairing its ability to learn from other features. Normalizing the data to a common scale allows the neural network to train more efficiently and converge to a solution more quickly, leading to improved model performance.

3.2 Analysis of the Results

The detailed examination of our training outcomes is visually represented in Fig. 3, which contrasts the machine learning performance metrics with the number of features, uncovers essential insights into our models' predictive capabilities. We emphasize the two top-performing models by marking their corresponding bars in red. Notably, the optimal performance is achieved when utilizing the five most important metrics: oxygen saturation (O2SAT) [31, 34, 37], partial pressure of carbon dioxide (PCO2) [13, 29, 30], age [10, 36], percentage of lymphocytes (Lymphocytes pct) [9, 22], and percentage of neutrophils (Neutrophils pct) [24, 26, 28].

However, it is important to note that if we only analyze the precision metric alone, the model incorporating these five key metrics ranks third. In this scenario, the model that employs only O2SAT as the input for training the neural network achieves 100% precision, accurately identifying all positive cases without any false positives. This metric is especially crucial in scenarios where false positives can lead to serious consequences, although it does not take false negatives into account. Despite this, for the other metrics evaluated, the model using the five most crucial biomarkers stands out as the best performing, reinforcing its effectiveness and the importance of integrating these five essential measures to improve model performance. This observation is critical as it demonstrates the advantage of a targeted approach compared to models incorporating a wider range of parameters.

The diminishing predictive power of models with an increasing number of features beyond the critical five metrics underscores a pivotal concept in machine learning: the paradox of complexity [19]. Incorporating more metrics does not necessarily enhance model performance; in fact, it can lead to increased complexity, diminishing returns, and potentially, a decrease in the model's ability to generalize to new data. This phenomenon can be attributed to several factors:

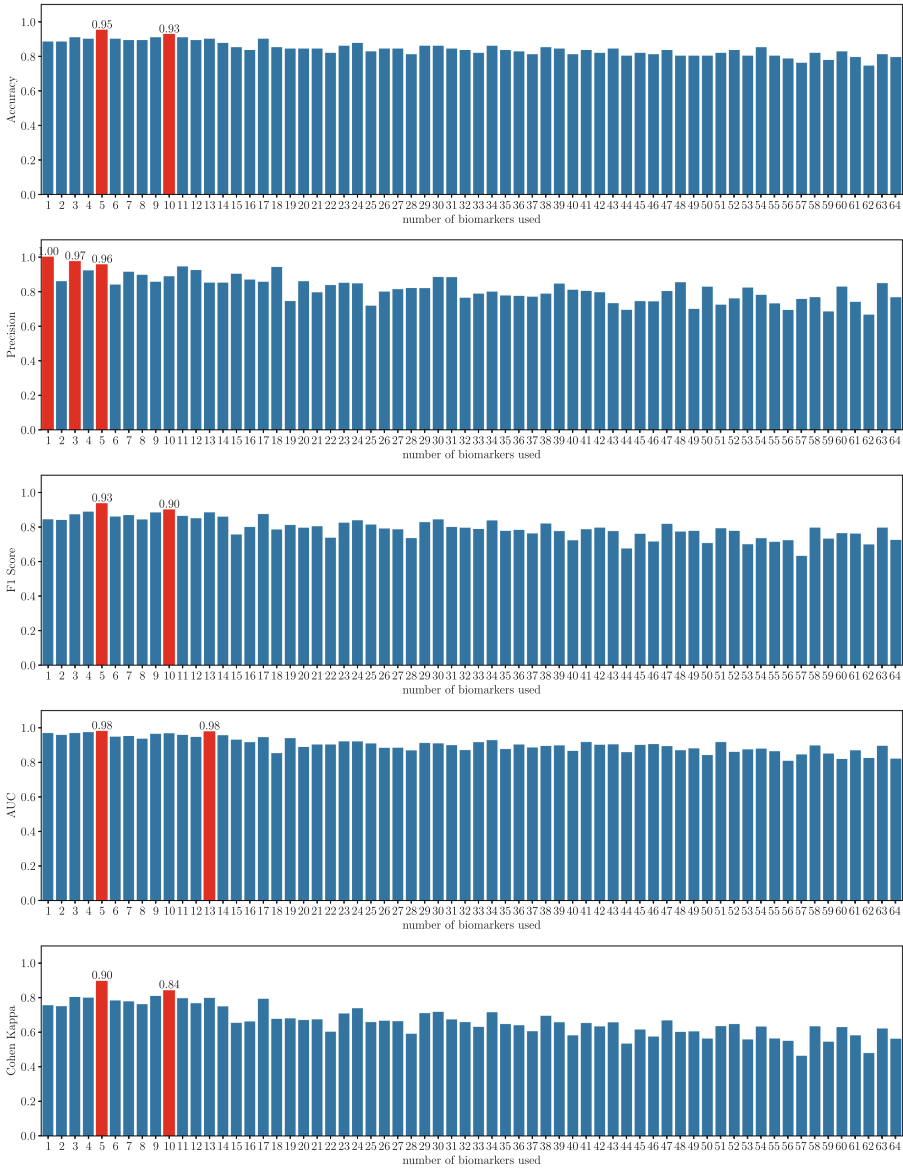


Fig. 3. Summary of Testing Results with Incremental Feature Inclusion

- A complex model with too many features may fit the training data too closely, capturing noise as if it were a significant pattern, which impairs its performance on unseen data [11].

- Each additional feature introduces a new dimension to the problem space. When many of these dimensions contribute little to no predictive value, they can dilute the impact of truly relevant features [8].
- More features necessitate more parameters to be estimated, increasing the computational complexity and resource demands of model training and inference [41].

The results suggest that more efficient and cost-effective predictive models can be developed by focusing on a subset of highly predictive indicators [33]. This streamlined approach not only enhances the accuracy of predictions but also optimizes resource use in clinical settings, especially under high-demand or limited resource scenarios. By identifying the most predictive features, we provide valuable direction for future research and the development of clinical tools to more effectively predict and manage COVID-19 severity.

The high accuracy achieved in this study can be attributed to the targeted selection of key biomarkers through L1 regularization and the optimization of neural network architectures using the Hyperband method. Unlike other models that can be present in the literature, which may utilize a broader range of features or different optimization techniques, our approach focuses on a minimal set of highly informative biomarkers, leading to enhanced model performance. While direct comparisons with other studies are challenging due to the unique characteristics of our dataset, the methodology employed here demonstrates superior efficiency and accuracy by leveraging only the most critical features, which is reflected in the high accuracy rates observed. Our findings advocate for a targeted approach in feature selection, emphasizing the importance of identifying and utilizing key metrics that offer the most significant predictive value. This strategy ensures the development of models that are not only accurate and efficient but also practical and interpretable within clinical and research settings.

4 Conclusions

We have demonstrated the potential of artificial intelligence in optimizing the prediction of disease severity, using COVID-19 as a case study. Our approach emphasizes the importance of feature selection and hyperparameter optimization. Employing L1 regularization and the Hyperband method, we identified a crucial set of biomarkers—oxygen saturation (O2SAT), partial pressure of carbon dioxide (PCO2), age, percentage of lymphocytes, and percentage of neutrophils—that together predict disease severity with remarkable accuracy.

Key conclusions from our research include: First, feature selection is essential in enhancing AI model effectiveness and efficiency in clinical settings, highlighting the importance of focusing on a select group of informative biomarkers. This approach could be particularly useful for addressing domain shifts or changes in data distributions, such as those caused by new viral strains like the JN.1 variant, which can impact predictive performance and accuracy. Second, the Hyperband method has proven effective for hyperparameter optimization in neural network

models, striking an optimal balance between predictive precision and computational complexity. Lastly, our findings provide promising avenues for using AI tools in the risk stratification of patients, facilitating more efficient resource allocation and improved patient management, potentially saving lives through early and personalized interventions.

This study is subject to several limitations. The reliance on a predefined tabular dataset constrained the ability to explore and extract novel features from alternative data types, such as radiological images or genomic sequences, which could further enhance predictive performance. While the methodology was developed with a focus on COVID-19, its principles are applicable to other diseases with similar data structures. However, comprehensive validation across diverse datasets is necessary to confirm its broader applicability. Despite these limitations, the framework developed in this study offers a robust foundation for future research, with opportunities for refinement through the integration of more complex data types and advanced AI models. Further research should also include a detailed evaluation of the clinical impact of these predictive tools in practical healthcare settings to optimize their utility.

References

1. Covid-19 coronavirus pandemic. <https://www.worldometers.info/coronavirus/> (2024), Último acceso: 12 de marzo de 2024
2. Alballa, N., Al-Turaiqi, I.: Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: a review. *Inform. Med. Unlock.* **24**, 100564 (2021). <https://doi.org/10.1016/j.imu.2021.100564>, epub 2021 Apr 3. PMID: 33842685; PMCID: PMC8018906
3. Aleem, A., Akbar Samad, A., Vaqar, S.: Emerging variants of sars-cov-2 and novel therapeutics against coronavirus (covid-19) (May 2023). <http://www.statpearls.com>, en: StatPearls [Internet]. Disponible desde: 2024 Jan
4. Appel, K.S., Geisler, R., Maier, D., Miljukov, O., Hopff, S.M., Vehreschild, J.J.: A Systematic Review of Predictor Composition, Outcomes, Risk of Bias, and Validation of Coronavirus Disease 2019 (COVID-19) Prognostic Scores. *Clinical Infectious Diseases* p. ciad618 (10 2023). <https://doi.org/10.1093/cid/ciad618>, <https://doi.org/10.1093/cid/ciad618>
5. Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. *Fut. Healthc. J.* **8**(2), e188 (2021)
6. Ballaz, S., Pulgar-Sánchez, M., Chamorro, K., Fernández-Moreira, E.: Scientific pertinence of developing machine learning technologies for the triage of covid-19 patients: a bibliometric analysis via scopus. *Inform. Med. Unlock.* **41**, 101312 (2023) <https://doi.org/10.1016/j.imu.2023.101312>, <https://doi.org/10.1016/j.imu.2023.101312>
7. Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.L., et al.: Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscipl. Rev. Data Min. Knowl. Discov.* **13**(2), e1484 (2023)
8. Brownlee, J.: Feature selection for machine learning in python. <https://machinelearningmastery.com/feature-selection-machine-learning-python/>. Accessed: 2024

9. Chen, G., Zhao, X., Chen, X., Liu, C.: Early decrease in blood lymphocyte count is associated with poor prognosis in covid-19 patients: a retrospective cohort study. *BMC Pulm. Med.* **23**(1), 453 (2023)
10. Chen, Y., Klein, S.L., Garibaldi, B.T., Li, H., Wu, C., Osevala, N.M., Li, T., Margolick, J.B., Pawelec, G., Leng, S.X.: Aging in covid-19: vulnerability, immunity and intervention. *Ageing Res. Rev.* **65**, 101205 (2021)
11. EliteDataScience: Overfitting in machine learning: what it is and how to prevent it. Accessed: 2024. <https://elitedatascience.com/overfitting-in-machine-learning>
12. Fine, M.J., Auble, T.E., Yealy, D.M., Hanusa, B.H., Weissfeld, L.A., Singer, D.E., Coley, C.M., Marrie, T.J., Kapoor, W.N.: A prediction rule to identify low-risk patients with community-acquired pneumonia. *N. Eng. J. Med.* **336**(4), 243–250 (1997)
13. Gounidis, A., Evangeliou, A.P., Kloura, C., Manganari, E., Parisi, C., Kourtidis, M., Kotronis, G., Apostolopoulou, M., Apostolidou-Kiouti, F.: Baseline hypocapnia is associated with intubation in covid-19 diagnosed patients, pp. 2021–11 (2021)
14. Hao, B., Sotudian, S., Wang, T., Xu, T., Hu, Y., Gaitanidis, A., Breen, K., Velmahos, G., Paschalidis, I.: Early prediction of level-of-care requirements in patients with covid-19. *Elife* **9**, e60519 (2020). <https://doi.org/10.7554/eLife.60519>. Oct
15. Hwangbo, S., Kim, Y., Lee, C., Lee, S., Oh, B., Moon, M.K., Kim, S.W., Park, T.: Machine learning models to predict the maximum severity of covid-19 based on initial hospitalization record. *Front. Public Health* **10**, 1007205 (2022)
16. Khosravi, M., Zare, Z., Mojtabaiean, S., Izadi, R.: Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health Serv. Res. Manage. Epidemiol.* **11** (2024). <https://doi.org/10.1177/23333928241234863>
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
18. Kistenev, Y.V., Vrazhnov, D.A., Shnaider, E.E., Zuhayri, H.: Predictive models for covid-19 detection using routine blood tests and machine learning. *Heliyon* **8**(10) (2022)
19. Krakauer, D.C., et al.: Unifying complexity science and machine learning. *Frontiers* (2023). <https://www.frontiersin.org/articles/10.3389/fams.2023.1061237/full>
20. Kwong, J.C., Nickel, G.C., Wang, S.C., Kvedar, J.C.: Integrating artificial intelligence into healthcare systems: more than just the algorithm. *NPJ Digit. Med.* **7**(1), 52 (2024)
21. Laatifi, M., Douzi, S., Bouklouz, A., Ezzine, H., Jaafari, J., Zaid, Y., El Ouahidi, B., Naciri, M.: Machine learning approaches in covid-19 severity risk prediction in morocco. *J. big Data* **9**(1), 5 (2022)
22. Lai, K.L., Hu, F.C., Wen, F.Y., Chen, J.J.: Lymphocyte count is a universal predictor of health outcomes in covid-19 patients before mass vaccination: A meta-analytical study. *J. Glob. Health* **12** (2022)
23. Levin, K.P., Hanusa, B.H., Rotondi, A., Singer, D.E., Coley, C.M., Marrie, T.J., Kapoor, W.N., Fine, M.J.: Arterial blood gas and pulse oximetry in initial management of patients with community-acquired pneumonia. *J. Gen. Intern. Med.* **16**(9), 590–598 (2001)
24. Li, J., Zhang, K., Zhang, y., Gu, Z., Huang, C.: Neutrophils in covid-19: recent insights and advances. *Virology* **20**(1), 169 (2023)
25. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: a novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**(185), 1–52 (2018)

26. Li, X., Liu, C., Mao, Z., Xiao, M., Wang, L., Qi, S., Zhou, F.: Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in covid-19 patients: a systematic review and meta-analysis. *Crit. Care* **24**(1), 1–10 (2020)
27. Lokhorst, J.: *The LASSO and Generalised Linear Models* (1999)
28. McKenna, E., Wubben, R., Isaza-Correa, J.M., Melo, A.M., Mhaonaigh, A.U., Conlon, N., O'Donnell, J.S., Ní Cheallaigh, C., Hurley, T., Stevenson, N.J., et al.: Neutrophils in covid-19: not innocent bystanders. *Front. Immunol.* **13**, 864387 (2022)
29. Ottestad, W., Søvik, S.: Covid-19 patients with respiratory failure: what can we learn from aviation medicine? *Br. J. Anaesth.* **125**(3), e280–e281 (2020)
30. Pulgar-Sánchez, M., Chamorro, K., Fors, M., Mora, F.X., Ramírez, H., Fernandez-Moreira, E., Ballaz, S.J.: Biomarkers of severe covid-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets. *Comput. Biol. Med.* **136**, 104738 (2021)
31. Qadir, F.I., Kakamad, F.H., Abdullah, I.Y., Abdulla, B.A., Mohammed, S.H., Salih, R.Q., Ali, R.K., Salh, A.M.: The relationship between ct severity infections and oxygen saturation in patients infected with covid-19, a cohort study. *Ann. Med. Surg.* **76** (2022)
32. Rikan, S.B., Azar, A.S., Ghafari, A., Mohasefi, J.B., Pirnejad, H.: Covid-19 diagnosis from routine blood tests using artificial intelligence techniques. *Biomed. Signal Process. Control* **72**, 103263 (2022)
33. Roland, T., Böck, C., Tschoellitsch, T., Maletzky, A., Hochreiter, S., Meier, J., Klambauer, G.: Domain shifts in machine learning based covid-19 diagnosis from blood tests. *J. Med. Syst.* **46**(5), 23 (2022)
34. Shenoy, N., Luchtel, R., Gulani, P.: Considerations for target oxygen saturation in covid-19 patients: are we under-shooting? *BMC Med.* **18**, 1–6 (2020)
35. Singh, D., Singh, B.: Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **97**, 105524 (2020)
36. Starke, K.R., Reissig, D., Petereit-Haack, G., Schmauder, S., Nienhaus, A., Seidler, A.: The isolated effect of age on the risk of covid-19 severe outcomes: a systematic review with meta-analysis. *BMJ Glob. Health* **6**(12), e006434 (2021)
37. Susanti, E., Okstoria, M.R., Wijayanti, S., Damayanti, H., Putra, H.E., Chipojola, R., Rahman, M.F., Kurniasari, M.D., Huda, M.H.: Clinical determinants of oxygen saturation and length of hospitalisation of covid-19 patients: a cross-sectional study in indonesia. *Enfermeria Clinica* **33**, S38–S44 (2023)
38. Vaishya, R., Javaid, M., Khan, I.H., Haleem, A.: Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes Metabol. Syndr. Clin. Res. Rev.* **14**(4), 337–339 (2020)
39. Vidaurre, D., Bielza, C., Larranaga, P.: A survey of l1 regression. *Int. Stat. Rev.* **81**(3), 361–387 (2013)
40. Visalakshi, S., Radha, V.: A literature review of feature selection techniques and applications: Review of feature selection in data mining. In: *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–6. IEEE (2014)
41. Ziolkowski, P.: Computational complexity and its influence on predictive capabilities of machine learning models for concrete mix design. *Materials* **16**(17), 5956 (2023). <https://doi.org/10.3390/ma16175956>



Predictive Model Proposal in Telemetry Using Machine Learning Techniques to Anticipate Water Degradation in Aquaculture

Néstor Rafael Salinas-Buestán^(✉) , Francisco Alexander Zambrano-Varela ,
Ángel Iván Torres-Quijije , Diego Fernando Intriago-Rodríguez ,
and Diego Patricio Peña-Banegas 

Universidad Técnica Estatal de Quevedo, Quevedo, Los Ríos 120301, Ecuador
{nsalinasb, francisco.zambrano2018, atorres, dintriagor3,
dpena}@uteq.edu.ec

Abstract. The present research presents an innovative strategy to improve water quality management in aquaculture environments by employing machine learning methods. A telemetric system was used to collect and monitor environmental data such as the power of Hydrogen (pH), Temperature, Turbidity and Total Dissolved Solids (TDS). The centrepiece of this study is the development of a telemetry-based predictive model, which leverages this data to predict changes in water quality. Several machine learning algorithms were tested, including Random Forest, Gradient Boosting, Polynomial Regression, Linear Regression, and k-Nearest Neighbors. After extensive testing, it was determined that Random Forest stood out as the most effective algorithm, achieving an extraordinary accuracy of 0.999 with a training time of just 12 s. This model allows for the proactive detection of conditions that could lead to the degradation of the aquatic habitat, enabling early warning of possible incidents. By anticipating these events, aquaculture resource managers can take corrective action in a timely manner, reducing the risks associated with water degradation. The results obtained underline the effectiveness and feasibility of applying machine learning methods in water quality management in aquaculture environments, showing that the combination of telemetry and appropriate algorithms can offer accurate and rapid solutions to improve sustainability and efficiency in this field.

Keywords: Aquaculture · Data Science · IoT · Machine Learning · Predictive Models · Telemetry

1 Introduction

The aquaculture, a crucial sector in food production, faces significant challenges in maintaining water quality, an essential factor for the health of aquatic organisms and crop productivity (Amenyogbe 2023; Colombo and Turchini 2021). Despite the use of telemetry systems to monitor water parameters in real time, the lack of efficient predictive tools that take advantage of these data to prevent water degradation limits the proactive actions of producers, increasing risks and losses in the sector (Thai-Nghe et al. 2020; P et al. 2023).

This rapidly expanding sector satisfies a significant portion of the world's demand for fish products, including popular species such as tilapia (Nie and Hallerman 2021). Globally, tilapia production has seen significant growth, driven by increasing demand in both local and international markets, particularly in the United States, which is the largest importer of tilapia worldwide (Miao and Wang 2020; Dai et al. 2020). In Ecuador, the production of red tilapia is being evaluated as a promising aquaculture activity in the Morona Santiago province (Edison Carillo et al. 2023).

The Internet of Things (IoT) in aquaculture involves the implementation of interconnected technologies designed to optimize operations, enhance sustainability, and protect the environment, ultimately improving the profitability and efficiency of aquaculture production (Acar et al. 2019). Through sensors, automation and real-time monitoring, IoT enables more accurate and efficient management of culture systems, optimizing water quality, feed and environment for fish (Bachri 2023; Antonio et al. 2023). However, the growth of this sector brings inherent challenges, with water quality management being one of the most critical for the survival and welfare of aquaculture species (Rastegari et al. 2023).

This work focuses on the development of a predictive model using machine learning techniques applied to telemetry data to improve the prediction and early detection of water quality degradation in aquaculture. This is carried out by identifying key variables, evaluating predictive models and validating the model through testing in real aquaculture environments. Through this system, it is possible to carry out proactive actions recognizing the degree of importance of integrating advanced solutions such as machine learning in the management of the quality of the fish environment, presenting an innovative approach to significantly transform efficiency and sustainability in the aquaculture sector, in this case the study introduces an advanced predictive model, specifically adapted for Red Tilapia.

1.1 Related Work

In the project by (Margarita Díaz Orozco et al. n.d.), uses predictive models in aquaculture but differs in methodology and variables compared to the current research. While focuses on numerical models for Red Tilapia, the current study uses machine learning to evaluate water degradation, considering biological and environmental factors. AI's role in aquaculture is highlighted, particularly in growth, water quality, and disease diagnosis.

The research by (Stankovic et al. 2020) explores the use of Fourth Industrial Revolution (4RI) technologies, such as AI and big data, in the public sector for water and sanitation in Latin America and the Caribbean. This study contrasts with the current research, which applies machine learning in aquaculture for the care of Red Tilapia, focusing on the specific challenges and implementation of these technologies in different sectors.

The article by (Chen et al. 2021) focuses on the design of a recirculating aquaculture system using machine learning methods. This work is relevant to the research as both studies apply machine learning in aquaculture. While the cited research focuses on system design, this work focuses on water degradation prediction making use of an end-to-end IoT system.

These studies show different approaches in the application of predictive models in aquaculture. While one of them focuses on a numerical model to improve the profitability

of Red Tilapia, the present study employs machine learning techniques to predict water degradation, considering biological and environmental variables. In addition, research on the use of artificial intelligence in areas such as biomass growth and disease diagnosis complement this approach. On the other hand, although the use of 4RI technologies in the public sector, specifically in water and sanitation, is addressed in another study, its relevance lies in the adoption of advanced technologies in similar contexts.

2 Methodology

2.1 Types of Investigation and Methodologies Applied

This study aims to discover the current characteristics of aquaculture water quality and to understand the relationships between key variables. Water conditions in aquaculture systems were mapped using telemetric data, identifying key patterns, which corresponds to Descriptive Research. Then, causes and correlations were analyzed, such as the impact of nitrates and external factors on water quality, framed as Explanatory Research. Subsequently, models were trained to predict future water degradation, using current and past data, as part of the Predictive Investigation. Finally, a statistical analysis was carried out to evaluate the effectiveness of the predictive models and the implemented methodology, which reflects as Quantitative Research. Across the development of the research, an observation of telemetric variables in aquaculture systems was carried out. The analytical method made it possible to evaluate and compare various predictive models available in the literature. Applying the synthesis method, it was possible to integrate the telemetric data under dynamic conditions, providing a complete evaluation of the effectiveness and accuracy of the model, thus ensuring a robust validation of the predictive model.

2.2 Data Collection

Data collection was carried out using a telemetry system equipped with sensors to measure four key variables: pH, Temperature, Turbidity and TDS. These sensors provided real-time measurements, allowing constant monitoring of water quality in the aquaculture environment. Telemetry facilitated automated and accurate data collection, essential for the development and validation of the predictive model (see Fig. 1). This system shows how node sensors connect to a central module, which transmits data via MQTT protocol to a cloud platform. The data is then accessible through various devices, enabling real-time monitoring through Wi-Fi or Ethernet.

2.3 Data Processing

A detailed analysis of the data set was carried out using tools such as Microsoft Excel and Tableau Prep Builder to organize the preliminary data on the most influential variables that affect water quality in aquaculture systems. In Fig. 2, it is shown how the data collected from the telemetry system, including pH, TDS, temperature and turbidity, undergoes cleaning and preparation before being analyzed.

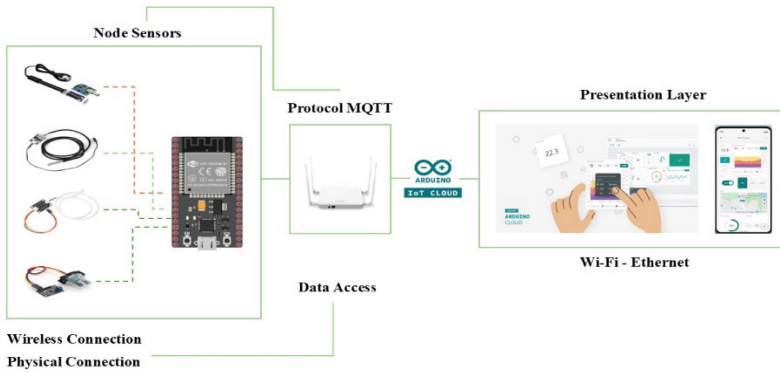


Fig. 1. A 3-layer telematic system architecture used to collect the dataset entries.

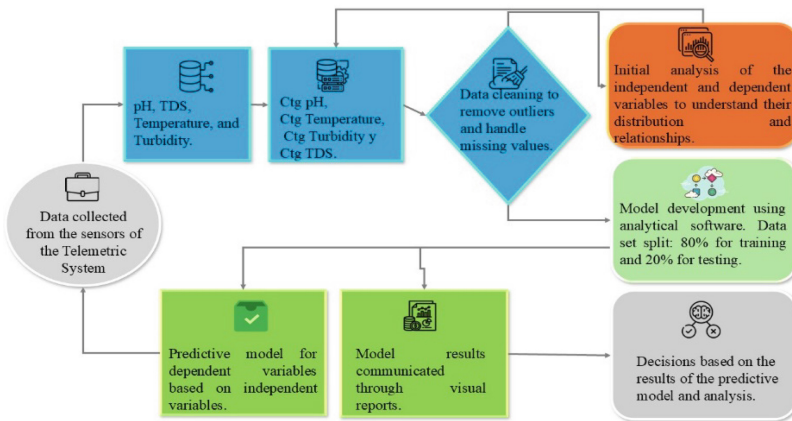


Fig. 2. Data processing flowchart: data analysis and model development.

2.4 Identification of the Variables Influencing Water Quality

Four key variables were identified to develop the predictive model. The independent variables selected are: ‘pH’, ‘Temperature’, ‘Turbidity’ and ‘TDS’. These variables are considered fundamental factors to anticipate and evaluate water degradation in aquaculture environments (Table 1).

Categorical variables were established by adjusting to discrete levels, these are: the ‘Ctg pH’, ‘Ctg Temperature’, ‘Ctg Turbidity’, ‘Ctg Total Dissolved Solids (TDS)’ and ‘Optimum Quality’. These categories are the values that allow prediction using the previously mentioned independent variables. These categories act as key indicators to evaluate water quality and contribute to the success of the predictive model in aquaculture (Table 2).

Table 1. Independent variables

Variable	Data type	Description	Min. value	Max. value
pH	Numeric	Acidity/Basicity	3	14
Temperature	Numeric	Degrees Celsius	4	37
Turbidity	Numeric	Water clarity	0	20.5
TDS	Numeric	Total dissolved solids	0	1000

Table 2. Dependent variables

Variable	Data type	Description	Min. value	Max. value
Ctg pH	Numeric	pH interval	1	5
Ctg temperature	Numeric	Temperature interval	1	5
Ctg turbidity	Numeric	Turbidity interval	1	3
Ctg TDS	Numeric	TDS interval	1	3
Optimum quality	Numeric	Water quality	0	1

2.5 Data Exploration

The unmodified dataset (rawdata) for this research was taken from the IoT system implemented for monitoring variables in aquaculture at the UTEQ university campus. The variables obtained are Temperature, pH, Turbidity and TDS. The Fig. 3 illustrates the dataset in its raw form, collected from the IoT system used for monitoring aquaculture variables before undergoing cleaning and preparation for further analysis.



Fig. 3. Telemetric system, located at UTEQ's La María Campus.

Red tilapia, a variety of tilapia with a distinctive color, are popular in aquaculture due to their rapid growth and resistance to various environmental conditions. The pond

has a system to monitor water status and control parameters such as water temperature, oxygenation, acidity and water quality.

As discussed in previous work by Díaz Orozco, this table details crucial influences of variables such as pH, Temperature, Turbidity, and TDS on aquaculture water quality. It establishes thresholds and effects, such as mortality, reproduction, and growth, providing guidelines for optimizing environmental conditions (Table 3).

Table 3. Critical variables and values range

Variable	Values
pH	Decreases pH (< 4.5): increases mortality 4.5 < pH < 6.5: decreases reproduction pH < 6.5: decreases growth 6.5 < pH < 8.5: optimum pH values 8.5 < pH: decreases growth 8.5 < pH: increase of non-ionized ammonium (toxic) 8.5 < pH: increase phytoplankton 8.5 < pH < 11: decreased reproduction 11 < pH < 14: increased mortality
Temperature	Decreases temperature (< 11 °C): increases mortality Decreases temperature: decreases metabolism and physiological processes Decrease in temperature (< 20 °C): decrease in reproduction Decreases temperature/Increases stress/Increases diseases/Increases mortality Optimal temperature values (22–32 °C): favors physiological performance, reproduction and avoids stress associated with extreme temperatures Increases temperature (> 37–38 °C): increases stress Increases temperature: decreases dissolved oxygen concentration
Turbidity	Increases turbidity: decreases alkalinity Increases turbidity: decreases growth Increases turbidity: decreases reproduction
TDS	Decreases TDS (< 200 ppm): reduction of the buffering capacity Optimal values: 200–500 ppm Increases TDS (> 500 ppm): increases fish stress/Decreases reproduction/Decreases growth

With a robust dataset consisting of 112,456 records for each variable, a valuable opportunity for detailed analysis and highly relevant predictive models could be performed. This amount of data, collected from June 2023 through December 2023, is a sample that effectively addresses the inherent variability, establishing a solid foundation (Table 4).

2.6 Correlation Matrix

This matrix allows identifying the significant relationships in the predictive telemetry model, focused on anticipating water degradation in aquaculture. It highlights the

Table 4. Descriptive statistics of the dependent variables used for the model generation.

Item	Ctg pH	Ctg TDS	Ctg Temp	Ctg Turb	Optimum quality
Count	112.546	112.546	112.546	112.546	112.546
Mean	3,403	1,644	3,657	1,312	0.1485
Std	0,877	0.688	0.9210	0.463	0.355
Min	1,000	1,000	1,000	1,000	0.000
25%	3,000	1,000	4,000	1,000	0.000
50%	3,000	2,000	4,000	1,000	0.000
75%	4,000	2,000	4,000	2,000	0.000
Max	5,000	3,000	5,000	2,000	1,000

strongest interconnections between the independent variables, such as pH, Temperature, Turbidity and TDS, and the dependent variables, such as Ctg pH, Ctg Temp, Ctg Turb, Ctg TDS, and the variable of interest Optimal Quality. It is noteworthy that, in this context, the tds, TDS Category and Turbidity variables show an outstanding correlation with the Optimal Water Quality variable. Figure 4 is presented below, showing the significant relationships between dependent and independent variables.

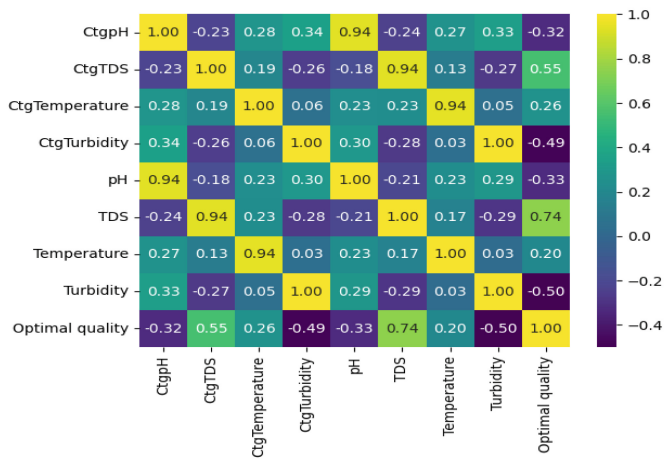


Fig. 4. The matrix reveals the connections between the water variables and the pond quality.

The values in the matrix vary from -1 to 1 , indicating the strength and direction of the correlation between the variables. A value of 1 indicates a perfect positive correlation, -1 a perfect negative correlation and 0 no correlation. Optimal quality has a strong positive correlation with “TDS” (0.74) and a negative correlation with “Turbidity” (-0.50).

2.7 Evaluation Between Different Prediction Models

Due to the nature of the data and the dependent variable, this research fits to supervised learning algorithms. Training tests were performed with the following algorithms: Gradient Boosting Regression, Random Forest, Linear Regression, Polynomial Regression and k-Nearest Neighbors. These algorithms allow the prediction of water degradation in aquaculture using advanced machine learning techniques. For the evaluation of their performance, the following metrics are considered: R^2 measures the proportion of variability explained by the model (value close to 1 indicates a good fit). MAE (Mean Absolute Error) is the average of the absolute errors between predictions and actual values. MSE (Mean Squared Error) is the average of the squared errors (penalizes larger errors). RMSE is the square root of MSE (provides error in the same units as the data).

To ensure the effectiveness of the predictive model, a data separation strategy was applied, which consists of using 80% of the data to adapt and train the model, reserving the remaining 20% for model validation and testing. This data partitioning enables the model to identify patterns and relationships from a large portion of the set, which contributes to its predictive capability.

Linear Regression. Linear Regression is a simple but effective method that establishes a linear relationship between the input variables and the output variable (Maydana Huanca 2021). In the case of anticipating water degradation, Linear Regression can provide a direct understanding of how variables affect water quality (see Fig. 5).

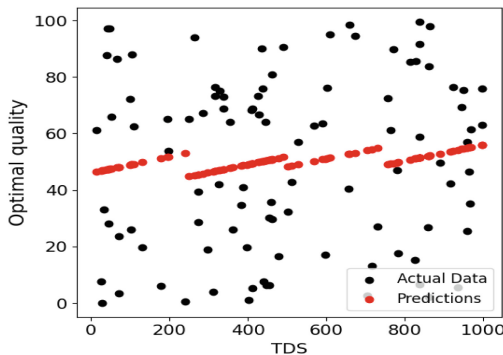


Fig. 5. Linear regression, optimal quality, as a function of TDS.

The result shows good overall performance with an R^2 of 0.80185 , indicating that it explains approximately 80% of the variability in the data. However, it has some shortcomings: an MSE of 0.049880 suggests that the line of best fit is far from many data

points, the RMSE of 0.22333 indicates small errors on average, but does not guarantee accuracy, and the MAE of 0.14548 reflects significant deviations in the predictions. Overall, although the model has acceptable explanatory power, its predictions may not be sufficiently accurate and may need improvement.

Gradient Boosting Regression. It is a supervised learning algorithm that combines multiple weaker models to form a robust and accurate model. It works by sequentially building decision trees, correcting for errors in the previous model at each step. This approach improves model accuracy and is effective in forecasting water degradation by identifying complex patterns in the data (Palaiokostas 2021).

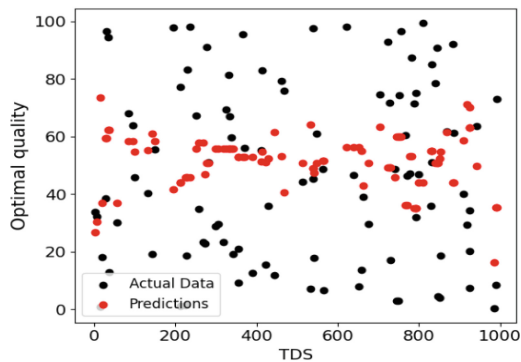


Fig. 6. Gradient Boosting Regression, optimal quality as a function of TDS.

The predictions of the Gradient Boosting Regression model (red dots) match the actual measurements (black dots) very well, reflected by an MSE of 0.0006380 and an RMSE of 0.0252600, indicating near-accurate accuracy (see Fig. 6). An R^2 of 0.9985448 confirms that the model predicts water quality from TDS with high accuracy. The high reliability of the model is demonstrated by its ability to explain almost all the variability in the observed data.

Random Forest. It is an algorithm that creates multiple decision trees and combines them to obtain a more accurate prediction that is resistant to overfitting (Swetha et al. 2023). In the context of aquaculture, Random Forest can identify complex relationships between variables to anticipate water degradation.

This model predicts “Optimal Quality” with high accuracy (see Fig. 7), as shown by an RMSE of 0.000006 and an R^2 of 0.999999, indicating that it explains 99.99% of the variability in pH-based water quality. The model predictions agree strongly with the actual values, and the MAE of 0.000006 suggests that the errors are minimal. The graphs visually reinforce the accuracy of the model, demonstrating that TDS levels are a strong predictor of water quality.

Polynomial Regression. It extends Linear Regression by introducing polynomial terms, allowing modeling of nonlinear relationships between variables. In aquaculture, where

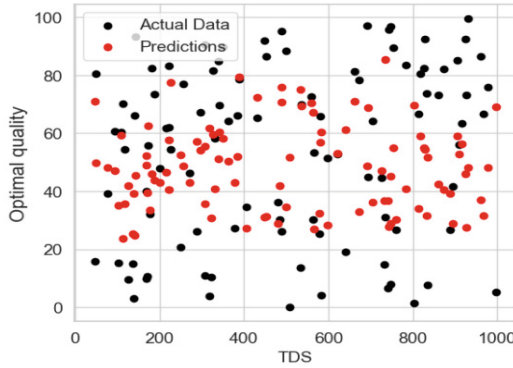


Fig. 7. Random Forest, optimal quality as a function of TDS.

interactions between factors can be more complex, it can be better when detecting variability in water degradation (Gómez Fernández 2022).

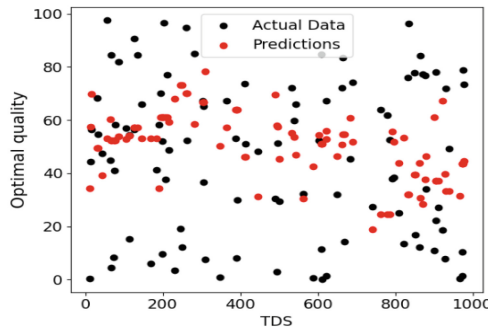


Fig. 8. Polynomial regression, optimal quality as a function of TDS.

The polynomial regression model created to approximate the relationship between TDS and water quality has an R^2 of 0.8766682, indicating that it explains 87% of the variability in the “Optimal Quality” of the water (see Fig. 8). Although this suggests a good fit, there is still 13% unexplained variability. The model shows that the amount of TDS significantly impacts water quality, suggesting that higher concentrations of TDS tend to negatively affect water quality.

k-Nearest Neighbors (k-NN). This algorithm is a supervised learning technique that can be applied in the aquaculture domain to predict water quality. In this approach, nearby points in the feature space are considered to share similar characteristics (Hamzaoui et al. 2023).

Evaluation of the k-Nearest Neighbors (KNN) model for predicting optimal water quality based on TDS shows an MSE of 0.002111 and an RMSE of 0.04595, indicating high prediction accuracy (see Fig. 9). The MAE of 0.00366 reinforces this accuracy.

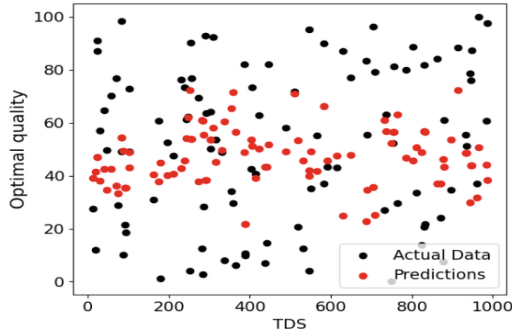


Fig. 9. k-Nearest Neighbors, optimal quality as a function of TDS.

An R2 of 0.9949 suggests that the model explains 99.49% of the variability in water quality, showing strong agreement between predictions and actual values. Overall, the KNN model proves to be very efficient and accurate in predicting water quality from TDS.

Table 5. Algorithm’s accuracy comparison results

Model	MSE	RMSE	MAE	R ²
Linear regression	0.04988	0.22333	0.14548	0.80185
Polynomial regression	0.03165	0.17791	0.10462	0.87666
Gradient Boosting	0.00063	0.02526	0.00117	0.99854
Random Forest	0.00000	0.00239	0.00002	0.99999
k-Nearest Neighbors	0.00211	0.04595	0.00366	0.99499

2.8 Computational Time and Selected Model

According to Table 5, it is evident that the model applying Random Forest offers the best fit to the aquaculture environment data. However, since it maintains a certain similarity, in terms of quadratic error, with Gradient Boosting, all models were also evaluated in terms of the time required for training. It can be seen in Fig. 10 that both models mentioned, present close levels, with a difference of 2 s.

Determination of an algorithm for a predictive model for aquaculture telemetry seeks to find the optimal balance between simplicity and capacity to capture the complexity of the environmental data, considering the computational variables. According to preliminary results, the model with the best quantifiable fit is based on the Random Forest algorithm, i.e., its use in aquaculture telemetry is supported by its exceptional capability, highlighted by a high coefficient of determination, reflecting its ability to outstandingly explain the variability present in complex aquaculture data. This adaptive capability is

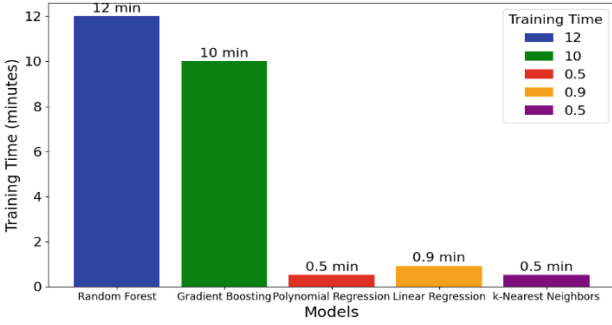


Fig. 10. Comparison of training ML models regarding the time required to train the dataset.

essential in aquaculture water quality management, where the simultaneous influence of multiple variables demands flexible and effective responses.

3 Results

To test the trained model, a web application has been created to allow aquaculture specialists to observe data reception in real time and perform a water quality prediction.

The web application makes real-time predictions using data transmitted from the telemetry system offering a valuable tool for aquaculture technicians.

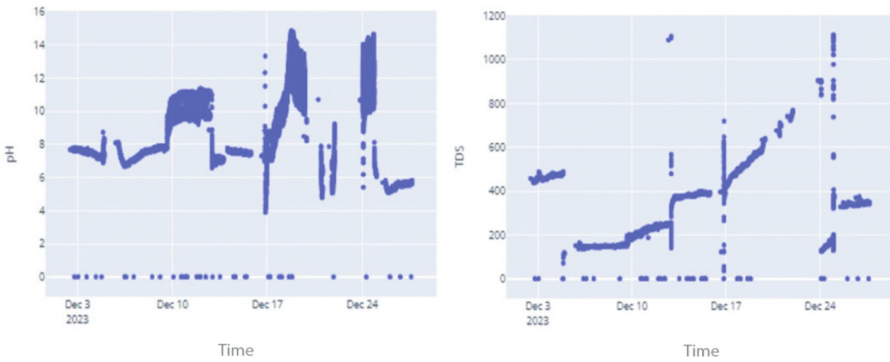


Fig. 11. Real-time pH and TDS predictions

The average values presented are pH at 7.44, TDS at 335, Temperature at 27.42, and Turbidity at 5.50 (see Fig. 11). Based on the interpretation of the prediction, pH is considered neutral, TDS moderate, Temperature high, and Turbidity low, resulting in an overall water quality rating of excellent (see Fig. 12).

At overall, this research determined the accuracy and effectiveness of several models using machine learning techniques. On the one hand, one of the predictive models achieves an exceptional accuracy of 99.9% with the Random Forest algorithm, which

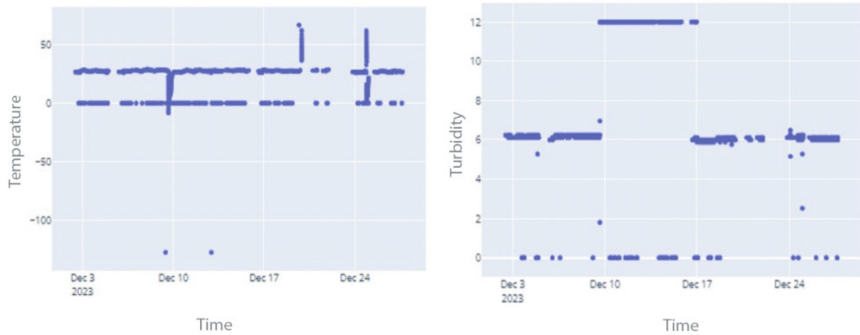


Fig. 12. Real-time temperature and turbidity predictions.

allows anticipating and solving environmental problems before they become worst. On the other hand, the article on “Artificial Intelligence in Aquaculture” also focuses on the application of artificial intelligence techniques to improve efficiency in aquaculture production, especially in the optimization of variables for growth and biomass. Although both articles share the goal of improving aquaculture through technology, the research stands out by prioritizing proactive water management, which makes it an essential element for long-term sustainability in this field.

The research also highlights its reliable tests in predicting four key variables using a telemetry system to collect and process data in real time. This ensures continuous and up-to-date monitoring of the aquatic environment, enabling accurate interpretations of water quality. In comparison to the article on 4RI technologies in water and sanitation, the research surpasses in providing a specific and highly effective method for anticipating and responding to changes in the environmental conditions.

4 Conclusions

The research highlights the relevance of data preprocessing in improving the quality and efficiency of predictive models. Careful data handling, including cleaning and proper partitioning between training and test sets, has been crucial to the success of the model. The importance of clean data is emphasized, as this allows accurate characterization of the performance and the relationship between variables in a holistic manner. Without clean data, there is a lack of a solid basis for further training or model choice.

The comparative study of several algorithms revealed that some algorithms, such as linear regression, were inadequate for this specific case, others, such as Random Forest and Gradient Boosting, exhibited high effectiveness in predicting water quality. Furthermore, the importance of statistical inference and correlation matrix analysis for an objective mathematical comparison of the algorithms’ performance is highlighted. This approach helps to express the results accurately through mathematical expressions, avoiding subjective interpretations in decision making.

Real tests in aquaculture are essential to validate the effectiveness of the predictive model based on telemetry and machine learning. Telemetry data collection is critical to

demonstrate the model's ability to predict water degradation in real aquaculture environments. These findings not only improve water management in aquaculture farms for a more sustainable operation, but also brings new research opportunities in the aquaculture field through the integration of IoT and Data Science technologies.




References

- Acar, U., et al.: Designing an IoT cloud solution for aquaculture. *Glob. IoT Summit (GIoTS)* **2019**, 1–6 (2019). <https://doi.org/10.1109/GIOTS.2019.8766428>
- Amenyogbe, E.: Application of probiotics for sustainable and environment-friendly aquaculture management—a review. *Cogent Food Agric.* **9**(1). (2023). <https://doi.org/10.1080/23311932.2023.2226425>
- Antonio, D.T.S., Rivera, J.R.A., Balid, A.C.N., Belaos, R.R.P., Brizuela, A.I., Caballero, J.A.: IoT-based water quality monitoring and automated fish feeder: enhancing aquaculture productivity. In: 2023 IEEE Region 10 Symposium (TENSYPMP), pp. 1–6 (2023). <https://doi.org/10.1109/TENSYPMP55890.2023.10223636>
- Bachri, A.: Freshwater monitoring system design in real-time for fish cultivation. *Int. J. Multi. Approach Res. Sci.* **2**(01), 362–371 (2023). <https://doi.org/10.59653/ijmars.v2i01.483>
- Carillo, E., Moncayo, Y., Rojas, L., Cárdenas, S.: Feasibility study for the implementation and definition of commercialization channels: red tilapia in the sucua canton province of morona santiago. *Russ. Law J.* **11**(7s) (2023). <https://doi.org/10.52783/rj.v11i7s.1073>
- Chen, F., et al.: Design of an intelligent variable-flow recirculating aquaculture system based on machine learning methods. *Appl. Sci.* **11**(14), 6546 (2021). <https://doi.org/10.3390/app11146546>
- Colombo, S.M., Turchini, G.M.: 'Aquafeed 3.0': creating a more resilient aquaculture industry with a circular bioeconomy framework. *Rev. Aquac.* **13**(3), 1156–1158 (2021). <https://doi.org/10.1111/raq.12567>
- Dai, Y.Y., Yuan, Y.M., Yuan, Y., Zhou, Z., Zhang, H.Y.: Competitiveness of Chinese and Indonesian tilapia exports in the US market. *Aquacult. Int.* **28**(2), 791–804 (2020). <https://doi.org/10.1007/s10499-019-00496-x>
- Gómez Fernández, D.: Análisis de la dinámica multitemporal del espejo de agua de dos lagos amazónicos durante 2014–2020, utilizando imágenes de SAR en GEE y técnicas de regresión (2022)
- Hamzaoui, M., Aoueleiyne, M.O.-E., Bouallegue, R.: A hybrid method of K-Nearest Neighbors with Decision Tree for water quality classification in aquaculture. In: International Conference on Computational Collective Intelligence, pp. 287–299 (2023)
- Margarita Díaz Orozco, M.C., Mario Gómez, J., Humberto Reyes, L.: Propuesta de un modelo predictivo numérico que permita mejorar la rentabilidad de la Tilapia Roja (n.d.)
- Maydana Huanca, A.R.: Elección del mejor modelo entre regresión lineal múltiple y árboles de regresión para predecir el precio máximo de las acciones de Intel en función al precio de apertura y volumen de ventas de acciones por día-2019 (2021)
- Miao, W., Wang, W.: Trends of aquaculture production and trade: Carp, tilapia, and shrimp. *Asian Fish. Sci.* **33**(S1), 1–10 (2020). <https://doi.org/10.33997/j.afs.2020.33.S1.001>
- Nie, P., Hallerman, E.: Advancing the sustainability of aquaculture. *Rev. Aquac.* **13**(2), 781–782 (2021). <https://doi.org/10.1111/raq.12548>
- Palaiokostas, C.: Predicting for disease resistance in aquaculture species using machine learning models. *Aquac. Rep.* **20**, 100660 (2021). <https://doi.org/10.1016/j.aqrep.2021.100660>

- Rastegari, H., Nadi, F., Lam, S.S., Ikhwanuddin, M., Kasan, N.A., Rahmat, R.F., Mahari, W.A.W.: Internet of Things in aquaculture: a review of the challenges and potential solutions based on current and future trends. *Smart Agric. Technol.* **4**, 100187 (2023). <https://doi.org/10.1016/j.atech.2023.100187>
- Stankovic, M., Hasanbeigi, A., Neftenov, N., Ventures, T.I., Basani, M., Núñez, A., Ortiz, R.: *Uso de tecnologías de la 4RI en agua y saneamiento en América Latina y el Caribe. El Agua y Saneamiento En América Latina y El Caribe* (2020)
- Swetha, P., Rasheed, A.H.K.P., Harigovindan, V.P.: Random Forest Regression based water quality prediction for smart aquaculture. In: *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pp. 1–5 (2023). <https://doi.org/10.1109/I3CS58314.2023.10127488>
- Thai-Nghe, N., Thanh-Hai, N., Chi, N.: Deep learning approach for forecasting water quality in IoT systems. *Int. J. Adv. Comput. Sci. Appl.* **11**(8) (2020). <https://doi.org/10.14569/IJACSA.2020.0110883>
- Vinoth Kumar, P., Suriya, K. Bala Murugan, D., Reshma, R.: Predicting and analyzing water quality using machine learning for smart aquaculture. In: *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pp. 354–359 (2023). <https://doi.org/10.1109/ICSCDS56580.2023.10104677>



Evaluation of the Use of Artificial Intelligence Techniques in the Mitigation of the Broadcast Storm Problem in FANET Networks

Andrés Sánchez^(✉) , Patricia Ludeña-González , and Katty Rohoden 

Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja, Loja, Ecuador
{adsanchez2,pjludena,karohoden}@utpl.edu.ec

Abstract. UAVs are nodes that can fly autonomously or can be operated remotely. The clustering of UAVs forms what is known as a FANET network. In a FANET the nodes mobility is much higher than in a MANET which results in more frequently changes in the network topology. Consequently, the broadcasting of packets are also expected to be executed more frequently which can cause redundancy, contention, and collision of packets, also known as the Broadcast Storm Problem (BSP). In this work, we propose the integration of artificial intelligence techniques to the analysis of the retransmission packets in a FANET. In this regard, various models commonly used in telecommunications networks were trained. As a result, an AI-based protocol that reduces the BSP problem is presented. Specifically, the AI-based protocol selects between two hybrid BSP reduction models. The results show that the proposal is meaningful in achieving greater efficiency in the retransmissions savings and thus in the energy consumption.

Keywords: Flying Ad-hoc Network · Broadcast Storm Problem · Machine learning techniques · Broadcast mitigation techniques

1 Introduction

The development of unmanned aerial vehicle (UAV) systems is the result of the rapid deployment of technological advances such as electronic, sensor, and communication technologies. UAVs are nodes that can fly autonomously or can be operated remotely. UAVs have been used mainly in military and civilian applications due to their flexibility, easy installation, and low operating costs [9, 12, 13, 15]. Nowadays, the proposal of using a group of small UAVs has called the attention of different applications since the many advantages that this presents. A Flying Ad-Hoc Network (FANET) has been defined as a network comprises of UAVs nodes. It is considered as a type of Mobile Ad-Hoc Network (MANET), where one of the main differences between them is the node mobility. In a FANET the nodes mobility is much higher than in a MANET which

results in more frequently changes in the network topology. Consequently, the broadcasting of packets are also expected to be executed more frequently which can cause redundancy, contention, and collision of packets, also known as the Broadcast Storm Problem (BSP).

The Broadcast Storm Problem has been widely studied in MANETs and VANETs. In [1], the authors describe in fully the IEEE 802.11 MAC protocol's operation, the system and communication models, and the formal specification of the broadcast operation for MANETs. It highlights the benefits of the alternating MultiPoint Relay (MPR) technique in balancing efficiency, effectiveness, and energy fairness while comparing other MANET broadcast algorithms. A Dynamic Broadcast Storm Mitigation Algorithm for VANETs is presented in [5]. This algorithm considers a safety following distance based on a 3s rule. In addition, the mobile velocity is used to compute a corresponding safety message broadcast time. However, in recent years, the solution of the BSP has been tackled in FANETS. The authors in [16] proved that when the number of UAVs increases, the BSP causes network contentions. Regarding this, authors proposed a dynamic neighborhood-based algorithm to mitigate the broadcast storm problem. Following this research line, in our previous work [17], some basic schemes and hybrid retransmission algorithms used to reduce the effects of BSP in MANET networks were adapted to FANET networks. The main contribution of this work was to determine the best scheme to avoid the BSP problem in FANET networks.

Lately, artificial intelligence has had a significant boom, especially in areas that involve troubleshooting in the operation and administration of networks. However, when integrating artificial intelligence techniques into the analysis of packet retransmission, some problems arise that require a solution of their own according to the technique being implemented. In addition, it is necessary to determine the parameters that directly and to a greater extent affect the decision to retransmit or not an incoming packet. This in order to improve the accuracy of the AI model when making the decision.

The main contributions of this work are:

- Review of AI techniques used in IoT networks.
- Performance evaluation of AI techniques applied to the BSP problem, for which various models popularly used in telecommunications networks were trained.
- An AI-based protocol that reduces the problem of BSP. This protocol uses AI to select between two hybrid BSP reduction models. Adapting the selection of the broadcast to the characteristics of the environment.

The remainder of this paper is organized as follows. In Sect. 2, we provide the related work. Section 3 presents the simulation scenarios, the simulation parameters, and the trained ML models. The results and the discussion of the results are described in Sects. 4 and 5. Finally, the conclusion is presented in Sect. 6.

2 Related Work

In most related works, the BSP problem is proposed to be solved by reducing redundant retransmissions. For example, in [8], a fuzzy logic-based routing scheme is presented to fly ad hoc networks. The routing scheme is comprised of two phases: the route discovery phase and the route maintenance phase. Authors state that the proposed technique in the first phase will be able to prevent the broadcast storm problem. However, there are no results that demonstrate this statement. In [18], authors use machine learning to detect protocol misbehaviour in VANETs. Results show that accurate and timely detection of protocol misconduct during emergency broadcasts significantly decreases the chances of broadcast storms.

In [3], the AID algorithm is analyzed, which is an adaptive packet retransmission algorithm integrated with an AI technique. This algorithm implements a decision tree model, that requires as input variables the nodes close to the emitter node. To obtain these data, it is necessary for the nodes to share their parameters (speed, packet counters, distance) with all their neighboring nodes so that the AI model can make the best decision. This exchange of information between nodes can cause an overload on the network.

In [11], the authors explore the application of AI techniques to packet retransmission algorithms, focusing on adapting to environmental conditions using a linear regression model. However, understanding the algorithms' workings is crucial for determining efficiency variables. Simulations with different settings are needed to enhance performance, incorporating information about the sending node's environment.

In order to determine whether to retransmit an incoming packet, in [21] authors investigate the usage of a convolutional neural network (CNN) model, which necessitates a huge training data set and network load. To ascertain the state and connection of the network, the model also necessitates computations and information sharing across nodes. AI methods can also be used to create the best retransmission algorithms.

Two AI methods are incorporated into the DBCC algorithm, which is examined in [10], to monitor and forecast link conditions between nodes and decide if an incoming packet has to be retransmitted. Support Vector Machines (SVM) are the first artificial intelligence (AI) approach it uses to forecast the connectivity conditions between nodes. Due to its sensitivity to noise in the input data, this SVM model requires a substantial amount of training data. A decision tree model is the second method that this program uses. The model's AI training requires extensive data, causing network overload and increasing computational costs. The data, primarily related to the node parameters, also consumes energy, highlighting the need for a more efficient network design.

In [19] a machine learning model is applied to a distance-based packet retransmission algorithm. For this algorithm, 3 variables are necessary: the mean distance between the emitting node and its neighboring nodes, the quadrant statistic and the Rician fading K factor. To obtain the first variable, it is necessary to exchange information (number of neighboring nodes, distances, quality of the

link) every certain period of time between the nodes, in order to keep the information updated. Once this information is obtained, the last two variables are calculated. The quadrant statistic allows one to measure how uniformly spaced a set of points are, and thus to obtain the mean distance between the neighboring nodes.

3 System Model

A node is an UAV that has a limited transmission range (see Fig. 1(a)). To maintain coordination, UAVs must relay messages across the network continuously, as can be seen in Fig. 1(b). As stated before, this message exchange can cause the broadcast storm problem. Thus, the main goal of our work is to determine the retransmission algorithm that reduces the maximum number of retransmitted messages based on artificial intelligence. In [4], three approaches are studied to integrate ML models in MANET networks and reduce the effects of BSP. The first of them focuses on integrating ML models to determine if the node should retransmit the packet. For this, some node parameters are selected (node speed, number of neighboring nodes, number of duplicate packets, etc.) and thresholds are defined for the selected parameters. Based on the selected parameters and their respective thresholds, the model must decide whether the packet should be retransmitted.

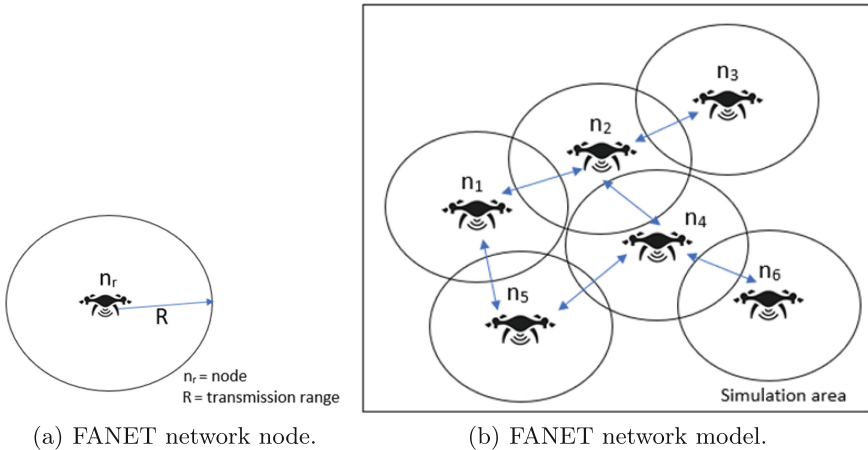


Fig. 1. FANET network structure.

The second approach incorporates ML models to improve the performance of the retransmission algorithms. In this case, the model is responsible of calculating the value of the algorithm’s decision parameter based on the network and the node parameters. After the ML model calculates the value of the decision parameter, the algorithm decides to retransmit the packet.

The third approach uses ML models to decide the retransmission algorithm to run. It is necessary to previously choose two packet retransmission algorithms and select parameters that allow us to analyze the performance of the selected algorithms. ML models are trained so that, based on the selected performance parameters, the algorithm that presents the best performance in that network scenario is chosen and executed to retransmit the packet. This last approach is studied in the present work in a FANET network scenario.

3.1 Generating Datasets

The selected packet retransmission algorithms were an algorithm based on probability and counter and an algorithm based on distance and probability. These algorithms are of the hybrid type (consisting of two basic algorithms) and are described in more detail in [17]. In order to train the ML models, the necessary datasets were generated using the previously indicated algorithms. Therefore, it is necessary to perform multiple simulations, extract the data generated and save them in a file. The simulation parameters for generating the datasets must be the same for both retransmission algorithms. This is necessary to compare the performance of both algorithms in each network scenario.

We consider a simulation area that ranges from 500 to 2500 m^2 , which increases in 100 m^2 to evaluate the effect of increasing the simulation area on the performance of the algorithms. Within this area several UAV nodes are deployed as shown in Fig. 1. Let $N = \{n_1, n_2, \dots, n_r\}$ be the set of UAV nodes. In addition, the amount of nodes, the transmission range, and the mobility algorithm are fixed for all the simulations. 693,178 simulations were carried out for each retransmission algorithm and from each one of them a file was extracted which contains the generated dataset, giving a total of 1386 generated datasets. All simulations were performed in a total of 10 attempts analyzing the range of number of nodes present in the network (50–150).

The communication parameters between drones considered are the 2.45 GHz operating frequency, 2 dBi antenna gain, omnidirectional antenna, reception threshold of -110 dB and an EIRP of 100 W. In addition, an obstacle-free environment was considered and the drones have LOS with their neighbors. In Table 1 the simulation parameters for the generation of the datasets are summarized.

According to [17], the parameters that determine the performance of the algorithms are the range and the saved retransmissions. The decision parameters that have to be defined are the input variables of the ML model. For this work, the chosen decision variables were the number of neighboring nodes and the speed of the nodes.

To choose the algorithm that has the best performance in each scenario, the percentage of reach achieved was taken as a basis. If both algorithms present the same range percentage, the algorithm that generates the highest percentage of retransmissions saved will be chosen. Thus, the training variable was elaborated by extracting the number of neighboring nodes, the average speed of the nodes, and the retransmission algorithm that presented the best results.

Table 1. Simulation parameters to generate the dataset.

Parameters	Values	Parameters	Values
Simulation area (m ²)	500–2500	Package size (bytes)	128
Number of nodes	50	Bandwidth (Mbps)	3
Nodes speed (Km/h)	0–110	Number of repetitions	10
Transmission range (m)	500	Mobility algorithm	Random walk

3.2 Training Machine Learning Models

This work was developed using the Classification Learner (CL) app of the machine learning and deep learning toolbox of the MATLAB software. This app has some models to train and are detailed in [14]. To train the available ML models, certain steps must be followed, such as the training variable must be loaded in the MATLAB workspace; this variable has to be a categorical type since it contains numerical values (number of neighbouring nodes and average speed of the nodes) associated with character strings (retransmission algorithm). Then, CL app automatically recognizes the input data (number of nodes and speed) and the output data (retransmission algorithm). If the variable is not of a categorical type, the input and output data of the ML model must be manually selected, in order to avoid possible errors when performing the training of the models. Additionally, the validation type Holdout at 20% is selected. This option causes a certain percentage of the training variable data to be used to check the precision of the trained model.

After loading the training variable into the app, all available models can be trained in one session. The time it takes for each model to perform its training depends on the complexity of the model and the size of the training variable. If the training variable has many input values, the model will take longer to train. Once the training process of the selected models is finished, they must be exported to the workspace to be saved in a file, then, they can be integrated into the FANET network scenarios to analyze their performance.

3.3 Mobility Model

There are some mobility models that can be integrated to simulate the movements of nodes within the network. These models are divided into four groups: purely random, time-dependent, planned routes and group mobility models; these are described in [2, 6, 7, 20].

The mobility model integrated in the present work is the Random Walk. This model allows to simulate a more dynamic network, since the nodes of the network are in charge of choosing themselves the speed and direction in which to move. That is why this model is one of the most used to simulate the movement of high mobility networks such as a FANET network.

4 Proposal

This section describes the packet retransmission algorithms selected to be integrated into the ML models, as well as the trained ML models, and the percentage of precision achieved by each of them.

4.1 Retransmission Algorithms

The retransmission algorithms selected to implement in this work are hybrid models, which are made up of two basic packet retransmission schemes. The first one is made up of the schemes based on counter and probability. It combines the performance of the basic counter-based scheme with a variant of the probabilistic scheme in which the threshold probability value is calculated based on certain environmental characteristics of the node. It is called a dynamic probability plus counter.

The second algorithm considers the basic schemes based on distance and probability. Like the previous algorithm, this one integrates a variant of the probabilistic scheme, in which the packet retransmission probability threshold value is calculated based on parameters of the sending node environment. It is called dynamic probability plus distance. The details of both algorithms can be found in [17]. For the generation of the datasets, these algorithms were integrated into the simulation parameters proposed in Table 1.

4.2 Model Training

For the present work, all ML models available in the MATLAB Machine Learning and Deep Learning app were trained. This in order to evaluate their performance and choose those who present the best results to implement them in the test scenarios. A total of 24 ML models were trained, the percentage of precision achieved by each of them can be seen in Table 2.

From each of the available models, a classifier type was selected to implement in the test scenarios. The classifiers chosen were those that presented the highest percentage of precision in each model. The selected classifiers were: Fine tree, quadratic discriminant, Naive Bayes kernel, quadratic SVM, medium KNN, bagged trees and the logistic regression model.

To obtain the results shown, around 600 simulations were performed for each of the ML models analyzed. These simulations were distributed so that each proposed scenario (1×1 , 3×3 , etc.) has at least 100 times, with different initial distributions and speeds in the nodes.

For the implementation of the proposed model in drones, the complexity of the model and the memory available in hardware must be considered. Models such as SVM and Ensemble are highly complex and require greater processing capacity and memory. On the other hand, models such as Decision Trees, Naive Bayes or Logistic Regression are the most suitable for implementation with drones, since these models do not require large memory spaces to store them and predictions are generated quickly, even on limited hardware.

Table 2. Precision of trained models.

Models	Types	Precision (%)
Decision trees	Coarse tree	99.3
	Medium tree	99.3
	Fine tree	99.3
Discriminant analysis	Linear	95.3
	Quadratic	98.0
Naive Bayes classifiers	Gaussian	97.3
	Kernel	98.6
SVM classifiers	Linear	99.3
	Quadratic	99.3
	Cubic	95.3
	Fine Gaussian	98.0
	Coarse Gaussian	95.9
KNN classifiers	Fine	94.6
	Medium	97.3
	Coarse	93.2
	Cosine	75.0
	Cubic	96.6
	Weighed	95.9
Ensemble classifiers	Boosted trees	70.3
	Bagged trees	98.6
	Subspace discriminant	70.3
	Subspace KNN	78.4
	RUSBoost trees	98.6
Regression	Logistic	99.3

5 Results

This section presents the results obtained by integrating the ML models in the FANET network scenarios. A performance comparison is also presented between having a network that integrates only relay algorithms and a network that integrates an ML model.

5.1 Models Performance

Figure 2 shows the percentage of reachability that is achieved with each of the ML models. All the models allow to achieve the maximum range in the first three scenarios, while in the last three scenarios it decreases, because the density of nodes is low and the nodes are very far apart and do not receive the packet, which reduces the range of the network. The decision tree model is the one that

allows a higher percentage of reach to be achieved, since up to the 7×7 scenario a reach close to 100% can be achieved. While the KNN model is the one that generates the least range, since between the 5×5 and 7×7 scenarios there is a difference of about 30% of range.

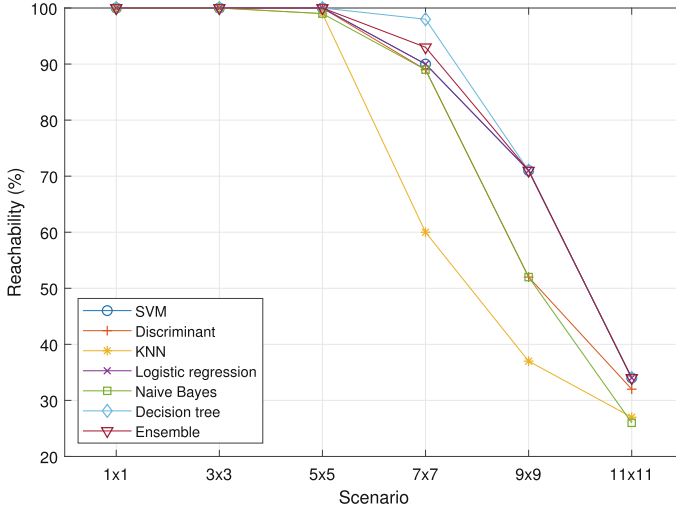


Fig. 2. Reachability achieved by each trained model.

The percentage of retransmissions saved is shown in Fig. 3. It is observed that all ML models achieve the same number of retransmissions saved in the first two scenarios. With the exception of the SVM model, which presents a difference of 15% in the first scenario. In the last three scenarios, the KNN model is the one with the lowest percentage of retransmissions saved. Since this model is the one with the lowest percentage of reach and, therefore, the saving of retransmissions is lower.

In the analysis of the power saving, it was decided to group the simulation scenarios into three groups. The first one made up of the 1×1 and 3×3 scenarios called the high node density scenario. The second brings together the 5×5 and 7×7 scenarios and is called the node-dense mean scenario. And finally the 9×9 and 11×11 scenarios form the low node density scenario. Figure 4 shows the power savings achieved by each of the ML models. It can be seen that in medium-density scenarios the greatest energy savings are generated, with the decision tree model being the one that generates the most energy savings. In scenarios of low node density, the KNN model is the one that generates the least energy savings because this model has the lowest percentage of range and the least amount of retransmissions saved. In high node density scenarios, all models generate the same power savings, except for the SVM model, since according to Fig. 3 this model presents 15% fewer retransmissions saved. Table 3 summarizes the amount of energy saved by ML models in each of the network scenarios.

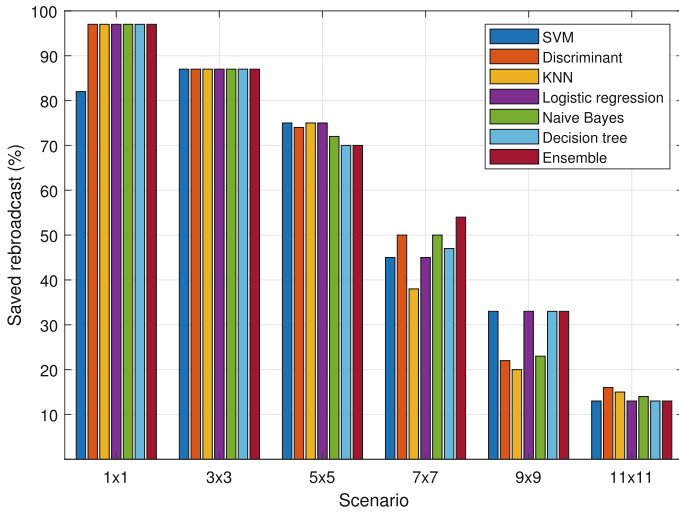


Fig. 3. Saved rebroadcast achieved by each trained model.

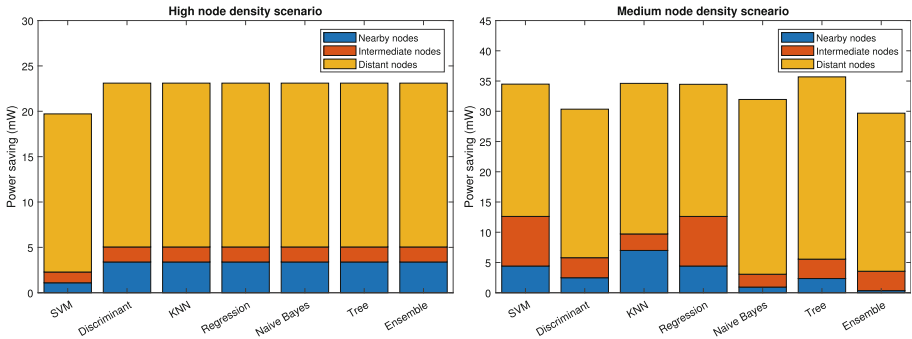
Table 3. Power saving achieved [mW] by each trained model.

Model	High node density	Medium node density	Low node density
SVM	19.71	34.49	10.37
Discriminant analysis	23.10	30.35	9.18
KNN	23.10	34.61	6.57
Logistic regression	23.10	34.46	10.37
Naive Bayes	23.10	31.95	9.68
Decision trees	23.10	34.51	10.37
Ensemble	23.10	29.69	10.37

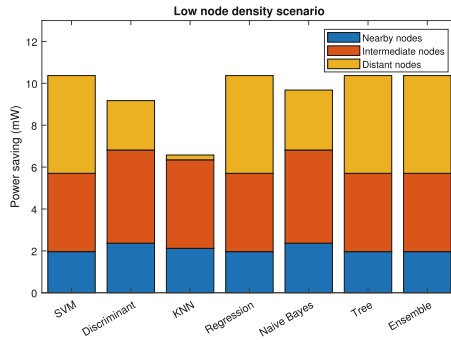
Table 4 shows the time required for each model to make a prediction in each network scenario. The high node density scenarios have the highest prediction times (on average 115 ms for the 1×1 scenario and 40 ms for the 3×3 scenario). In the scenarios of medium and low density of nodes, the lowest prediction times are presented. The Ensemble model takes the most time to make a prediction. The decision tree model is the one that generates the smallest prediction times in all network scenarios.

5.2 Algorithms Comparison

To determine the efficiency of integrating AI models in packet retransmission in a FANET network scenario, the model that presented the best results was selected and compared with the results of the hybrid algorithms obtained in [17]. The Decision Tree classifier has the highest percentage of reach and retransmissions



(a) Power saving in high node density scenario. (b) Power saving in medium node density scenario.



(c) Power saving in low node density scenario.

Fig. 4. Power saving achieved by each trained model.

saved of all the models analyzed, as well as the amount of energy saved. It also has the lowest execution times, making it the model that requires the least time to make a prediction; for these reasons, it is the selected algorithm.

Then, a comparison was made to analyze the efficiency when the AI model is integrated with the network and the efficiency without the AI model. Figure 5 shows the percentage of range obtained by the hybrid algorithms, where Hybrid 1 is the algorithm based on counter + probability and Hybrid 2 is the algorithm based on distance + probability, the model of Selected AI and a network model that only uses Flooding as a packet relay method. It can be seen that Flooding is the one that generates the most reach in the network, followed by the AI model. The hybrid algorithm based on counter probability + is the one that generates the lowest percentage of reach in the last scenarios.

Figure 5 shows the percentage of retransmissions saved by hybrid algorithms and the AI model. The Flooding method is excluded because it does not generate retransmission savings. In the first three scenarios, hybrid algorithm 2 is the one that generates the lowest percentage of retransmission savings, while hybrid

Table 4. Prediction time achieved [ms] by each trained model.

Model	1 × 1	3 × 3	5 × 5	7 × 7	9 × 9	11 × 11
SVM	121.71	29.40	3.16	3.19	2.90	2.90
Discriminant	100.33	32.34	4.01	3.66	3.99	3.96
KNN	122.13	39.30	3.92	4.28	4.12	4.33
Regression	152.73	40.63	4.32	5.25	4.54	4.37
Naive Bayes	125.43	39.93	19.56	4.26	4.33	4.28
Decision tree	84.97	27.80	3.10	2.89	2.89	2.91
Ensemble	150.20	68.17	34.92	38.25	35.37	35.51

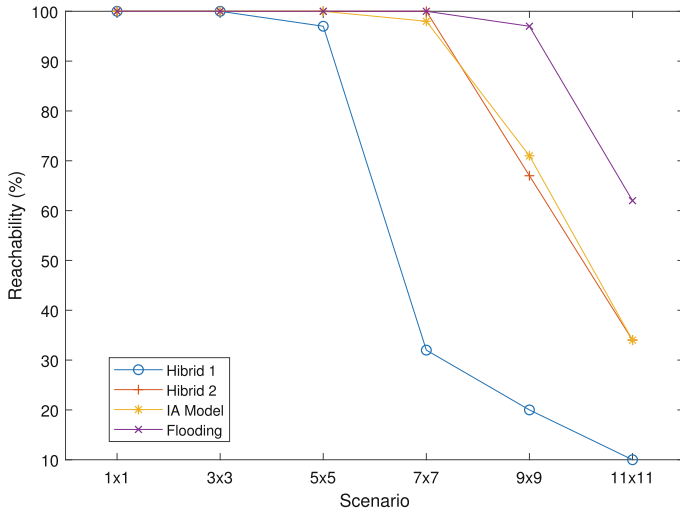


Fig. 5. Reachability comparison between hybrid algorithms, IA model and flooding.

algorithm 1 is the one that achieves the most retransmission savings, followed by the AI model (with a difference between 1 and 2% between both). In the last three scenarios, the AI model is the one that generates the highest percentage of retransmission savings, while hybrid algorithm 1 presents the lowest retransmission savings.

The comparison of the amount of power saved is shown in Fig. 7. It is observed that the hybrid algorithm 1 is the one that saves the most energy, followed by the AI model and the hybrid algorithm 2. The AI model generates improvements in the performance of the network compared to having only the hybrid algorithms integrated in the network. A greater range is achieved in the network, maintaining a high percentage of retransmission savings. And it is possible to balance the power saving between the nodes (near, intermediate and far) of the network, as shown in Fig. 7.

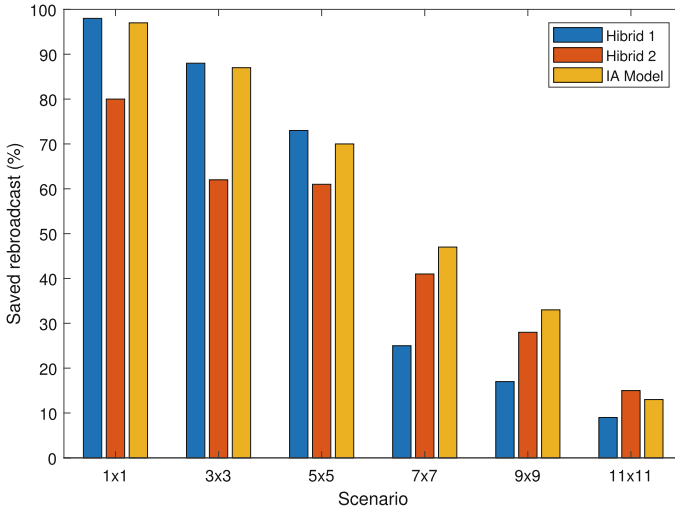
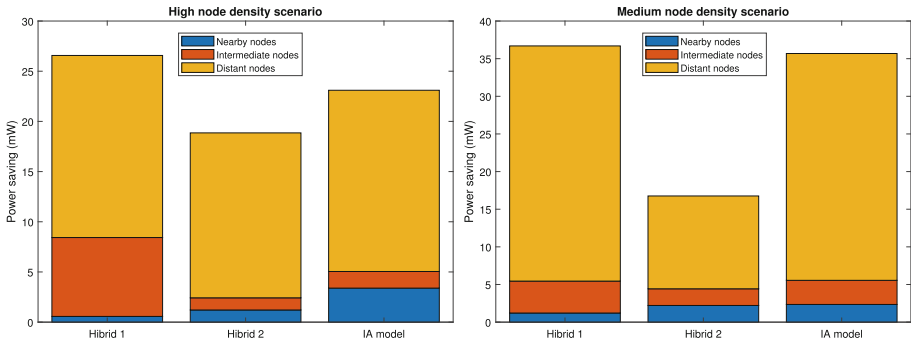
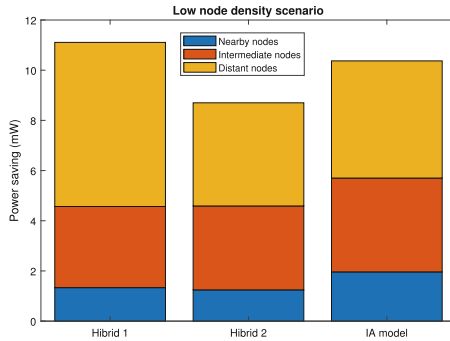


Fig. 6. Saved rebroadcast comparison.



(a) Power saving in high node density scenario. (b) Power saving in medium node density scenario.



(c) Power saving in low node density scenario.

Fig. 7. Power saving comparison.

6 Conclusions

The integration of AI techniques to control the BSP problem is an innovative approach that opens a new line of research. In this work, a literature review was carried out on the integration of AI models in FANET networks, in this review it was observed that there are not many related works developed, giving more importance to the present work and future developments on this topic. In addition, the approach to follow for the development of this work was defined and the packet retransmission algorithms to be evaluated and implemented in the AI models were selected.

To evaluate the proposal, six FANET network scenarios were defined and their performance based on evaluation metrics. The results show that the proposal is valid and that they manage to be more efficient in saving retransmissions and therefore in energy consumption. The model that generates the highest percentage of reach is the decision tree model, since until the 7×7 scenario it presents values close to 100%. However, the KNN model is the one that generates the lowest percentage of reach in the last three scenarios. In addition, all the models present similar percentages of saved retransmissions, except for the SVM model since, in the 1×1 scenario, it presents 20% less than the other models. While the KNN model presents a lower percentage of retransmissions saved in the last scenarios because its reach percentage decreases in these scenarios.

Regarding the execution times of the AI models, it was found that they are all similar in all network scenarios. In the case of the 1×1 scenario, execution times between 7 and 10 ms occur due to the high density of nodes. In the low-density scenarios, the execution times are close to 0.2 ms. In conclusion, the decision tree model performs best when implemented in an FANET network model, as it generates the highest reach and retransmission percentages saved, and the lowest execution times and latency.






References

1. Agon-Rambosson, A., Lejeune, J., Sopena, J., Sens, P.: Alternating MPR: a balanced broadcast algorithm for Manets. In: 2022 IEEE 21st International Symposium on Network Computing and Applications (NCA), pp. 19–26. IEEE (2022). <https://doi.org/10.1109/NCA57778.2022.10013596>, <https://hal.sorbonne-universite.fr/hal-03967460>
2. Agrawal, J., Kapoor, M.: A comparative study of mobility models for flying ad hoc networks. *Int. J. Sens. Netw.* **38**(3), 204–214 (2022)
3. Bakhouya, M., Gaber, J., Lorenz, P.: Energy evaluation of AID protocol in mobile ad hoc networks. *J. Netw. Comput. Appl.* **58**, 287–293 (2015)
4. Colagrosso, M.: Intelligent Broadcasting in Mobile Ad Hoc Networks: Three Classes of Adaptive Protocols (2007). <https://doi.org/10.1155/2007/10216>
5. Feukeu, E., Zuva, T.: Dynamic broadcast storm mitigation approach for Vanets. *Futur. Gener. Comput. Syst.* **107**, 1097–1104 (2020)
6. Hassan, M.A., Imad, M., Hassan, T., Ullah, F., Ahmad, S.: Impact of routing techniques and mobility models on flying ad hoc networks. In: Computational Intelligence for Unmanned Aerial Vehicles Communication Networks, pp. 111–129. Springer (2022)

7. Kumari, K., Sah, B., Maakar, S.: A survey: different mobility model for Fanet. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **5**(6) (2015)
8. Lee, S.W., Ali, S., Yousefpoor, M.S., Yousefpoor, E., Lalbakhsh, P., Javaheri, D., Rahmani, A.M., Hosseinzadeh, M.: An energy-aware and predictive fuzzy logic-based routing scheme in flying ad hoc networks (FANETs). *IEEE Access* **1** (2021). <https://doi.org/10.1109/ACCESS.2021.3111444>
9. Li, Y., Bi, Y., Wang, J., et al.: Unmanned aerial vehicle assisted communication: applications, challenges, and future outlook. *Clust. Comput.* (2024). <https://doi.org/10.1007/s10586-024-04631-z>
10. Lyu, F., Cheng, N., Zhou, H., Xu, W., Shi, W., Chen, J., Li, M.: DBCC: leveraging link perception for distributed beacon congestion control in VANETs. *IEEE Internet Things J.* **5**(6), 4237–4249 (2018)
11. Mahajan, S., HariKrishnan, R., Kotecha, K.: Prediction of network traffic in wireless mesh networks using hybrid deep learning model. *IEEE Access* **10**, 7003–7015 (2022)
12. Mohsan, S.A.H., Othman, N.Q.H., Li, Y., Alsharif, M.H., Khan, M.A.: Unmanned aerial vehicles (UAVs): practical aspects, applications, open challenges, security issues, and future trends. *Intel. Serv. Robot.* **16**(1), 109–137 (2023)
13. Nguyen, M.T., Nguyen, C.V., Do, H.T., Hua, H.T., Tran, T.A., Nguyen, A.D., Ala, G., Viola, F.: UAV-assisted data collection in wireless sensor networks: a comprehensive survey. *Electronics* **10**(21) (2021). <https://doi.org/10.3390/electronics10212603>, <https://www.mdpi.com/2079-9292/10/21/2603>
14. Paluszek, M., Thomas, S., Paluszek, M., Thomas, S.: Matlab machine learning toolboxes. *Practical MATLAB Deep Learning: A Project-Based Approach*, pp 25–41 (2020)
15. Pasandideh, F., da Costa, J.P.J., Kunst, R., Islam, N., Hardjawana, W., Pignaton de Freitas, E.: A review of flying ad hoc networks: key characteristics, applications, and wireless technologies. *Remote Sens.* **14**(18), 4459 (2022)
16. Pires, R.M., Pinto, A.S.R., Branco, K.R.L.J.C.: The broadcast storm problem in fanets and the dynamic neighborhood-based algorithm as a counter measure. *IEEE Access* **7**, 59737–59757 (2019)
17. Sánchez, A., Ludeña-González, P., Sandoval, F., Rohoden, K.: Evaluation of the Broadcast Storm Problem Based on Hybrid Retransmissions Algorithms in Fanet Networks (2020)
18. Sharshembiev, K., Yoo, S.M., Elmahdi, E.: Protocol misbehavior detection framework using machine learning classification in vehicular ad hoc networks. *Wirel. Netw.* **27**, 2103–2118 (2021). <https://doi.org/10.1007/s11276-021-02565-7>
19. Tang, F., Mao, B., Kato, N., Gui, G.: Comprehensive survey on machine learning in vehicular network: technology, applications and challenges. *IEEE Commun. Surv. Tutor.* **23**(3), 2027–2057 (2021)
20. Wheeb, A.H., Nordin, R., Samah, A., Alsharif, M.H., Khan, M.A.: Topology-based routing protocols and mobility models for flying ad hoc networks: a contemporary review and future research directions. *Drones* **6**(1), 9 (2021)
21. Zhang, P., Gao, A.Y., Theel, O.: Less is more: learning more with concurrent transmissions for energy-efficient flooding. In: *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 323–332 (2017)



Adaptation Dynamics of Galápagos Finches: Evolutionary Responses to Climate Variation Explored through Machine Learning

Ariana Deyaneira Jiménez Narváez¹, Dánely Leonor Sánchez Vera¹,
Iván Reyes², Paulina Vizcaíno-Imacaña²,
and Manuel Eugenio Morochó-Cayamcela^{1,2}

¹ School of Mathematical and Computational Sciences, Yachay Tech University, Urcuquí 100119, Ecuador

ariana.jimenez@yachaytech.edu.ec, danelly.sanchez@yachaytech.edu.ec,
mmorochoco@yachaytech.edu.ec, mamorochoca@uide.edu.ec

² Faculty of Technical Sciences, School of Computer Science, Universidad Internacional del Ecuador, Quito 170411, Ecuador
ireyes@uide.edu.ec, pvizcaino@uide.edu.ec

Abstract. Climate change is accelerating the evolution of species as global warming intensifies and natural environments undergo significant alterations. This study addresses the urgent need to understand how these environmental changes impact avian populations, specifically focusing on the physiological adaptation of finch beaks on the Galápagos Islands. By employing multiple models-logistic regression, XGBOOST, and Random Forest Classifier-using the most recent data, the study investigates how changes in climate conditions correlate with variations in beak dimensions. The analysis reveals significant relationships between temperature, precipitation, and finch beak characteristics, indicating that climatic factors have a notable impact on beak morphology. The study aims to clarify the evolutionary dynamics at play and to predict future adaptations based on projected climate scenarios. The results highlight the robust correlation between climatic variables and finch physiology, providing deeper insights into the complex interplay between climate change and evolutionary biology. These findings are crucial for developing targeted conservation and adaptation strategies, helping to safeguard biodiversity in the face of ongoing climate shifts. The insights gained from this study are essential for guiding policymakers and stakeholders in implementing effective measures to mitigate the impacts of climate change on ecosystems and species survival.

Keywords: Evolutionary adaptation · Predictive modeling · Data analysis · Climate change · Artificial intelligence · Biodiversity Conservation

1 Introduction

The impact of climate change on biodiversity and agriculture has emerged as a critical area of research in recent years [1–4]. The finches of the Galápagos Islands, famously studied by Charles Darwin, serve as a valuable model for understanding how environmental factors drive evolutionary changes in species. The morphology of finch beaks, in particular, exhibits a high sensitivity to environmental variables such as temperature and precipitation, making them an ideal subject for examining the relationship between climate conditions and physical adaptations.

This study aims to predict future species distribution and morphological changes in Galápagos finches based on historical climate data and current beak measurements. Utilizing advanced machine learning algorithms, including Logistic Regression, XGBoost, and Random Forest Classifiers, this research analyzes the correlation between climatic variables and beak dimensions. These models facilitate informed predictions regarding the evolutionary trajectories of finch species under different climate scenarios.

The dataset from the year 2012, a period marked by notable climatic variability, serves as the foundation for training the predictive models. The efficacy of these models is assessed using accuracy metrics, and the most robust model is subsequently employed to forecast future species distributions. This approach not only enhances the understanding of the adaptive responses of finches to environmental changes but also contributes to broader conservation efforts by identifying potential future risks to biodiversity.

This investigation underscores the importance of integrating climate data with morphological measurements to predict species adaptations. The findings aim to provide valuable insights for ecologists, conservationists, and policymakers working to preserve the unique biodiversity of the Galápagos Islands amidst ongoing climate change.

2 State of the Art

The harmonious coexistence of nature has long confounded the minds of the thinkers of yesteryears. “The Origin of Species” by Darwin was perhaps the first scientific inquiry to unravel the elegant balance and collaboration between species that have been created to exist in perfect harmony in nature; a cohabitation that is possible only because of the evolution of morphological characters. The Galápagos Islands, notorious as the cradle of modern evolutionary thought, inspired this great naturalist Charles Robert Darwin into proposing the principles that guided evolution through the process of adaptation to the environment [5]. He remarked on the presence of finches in the islands and noted their phenomenal decoupling of size and shape of their beaks relative to location, which provided him with data for his next proposition: the principle of natural selection. These finches, situated in geographically close suites of islands, have evolved finely tuned morphological traits with respect to the environment where

they exclusively feed and thereby occupy discrete niches on the islands due to their differential utilization of food resources [6]. Because of the highly variable environment that the Galapagos offers, it made Darwin proposed that different species of finch had evolved from a common ancestor that had emigrated from mainland South America. Then, the impact of environmental changes, such as urbanization, changes too on the interactions of finches and other Galápagos species [7].

2.1 Historical Background of Finch Beak Studies

To explore confidence in the existence of true beak-shape evolution, hypothetical ancestor-descendant pairs representing directional selection were identified in a unique time-stamped finch beak database obtained from measurements made by Peter and Rosemary Grant and their colleagues [8]. The authors' fifteen-year tracking of the lineage pair from 1990 to 2006 offered the ability to monitor the direction of phenotypic variation, thereby suggesting a strong basis for inferences beyond the brief 'snapshot' counting of pandemonium that had fostered the evolutionary paradox [9]. Indeed, analysis of only a portion of the long dance of evolutionary life yielded a significant relationship to the evolutionary potential provided by the beak-shift experiment employing medium finches [10].

It was Charles Darwin's visit to the Galapagos Islands that first led to the theory of evolution, developed during his time in the islands, over the years 1831 to 1836 [10]. Darwin developed his theory of evolution through the observation of the many species of finches inhabiting the islands and their differing appearance, in particular their varying beak shapes [9]. The islands represented a unique laboratory of evolution that not only affected Darwin but generations of naturalists since him. Today, the Galapagos, especially the medium ground finch which comprises what are now designated as different species, share the focus of the international community because of the groundbreaking nature of the work despite the ongoing paradoxical and long-mystifying issue of apparently rapid evolution but lack of evolutionary conclusions [11].

Droughts and Beak Shifts The groundbreaking research of Peter and Rosemary Grant on the medium ground finch population on Daphne Major Island provides a striking illustration of how climate change can shape evolutionary trajectories. In the early 1980s, a series of severe droughts caused by El Niño events dramatically altered the island's ecosystem [12]. The availability of hard seeds, a crucial food source for medium ground finches, plummeted, leading to a period of intense selection pressure [13].

In response to this environmental shift, the finch population underwent a rapid and dramatic change in beak size. Birds with larger beaks, better equipped to crack open tougher seeds, gained a significant survival advantage. As a result, the average beak size in the population increased noticeably over a relatively short period [8]. This remarkable example of natural selection in action highlights the sensitivity of evolutionary processes to environmental perturbations.

2.2 Significance of Studying Finch Beak Evolution

The main reason for the lack of developmental studies on finch beak evolution is the limited availability of embryonic material for beak analysis. This is largely due to the strict conservation regulations imposed by the Galapagos National Park [14]. Thus, it has been decided to identify beak changes using an algorithm that first analyzes evolutionary responses according to climatic changes.

Galapagos finches are a popular model species for studying evolution because of their complex evolutionary history associated with rapid morphological divergence, geographic diversification, and ecological adaptation. Although Darwin considered finches to be strong evidence for his theory of evolution by natural selection, understanding of the developmental biology underlying their evolutionary transformation remains limited [12]. Of particular interest are the evolutionary forces underlying finch beaks, which have highly adaptive functions driven by different selection pressures on separate islands, especially now with the climatic changes we are experiencing [13].

2.3 Significance of Beak Morphology in Evolutionary Studies

The importance of beak morphology comes from the close relation of the form and function, as beak shape is expected to be evolutionarily shaped by its function or ecological adaptation. By adapting to their specific feeding tasks, the beak might develop into a certain length, width, and depth and adopt unique curves [15]. We can understand clear ecological relationships between beak morphology and foraging niches in different wild species, e.g., fruit seed consumers have thicker and narrower beaks, ground foragers have wider and deeper beaks, and wood-living invertebrate feeders have flatter, wider, and deeper beaks [14]. Researchers have used such relationships to solve past bird feeding adaptations. The structure-function correlation of beak shape results, because certain mechanical forces are required for the beak to perform necessary duties, and relationships to the muscular and skeletal system, suggest that evolutionary or developmental variations in beak size or shape are indicative of these types of adaptive pressures [10]. That is why beak shape diversity between species might be an evolutionary response to ecological factors in the wild [16].

As highlighted in the literature review, several studies have examined the evolution of finch beaks in the Galápagos Islands. However, the central focus of our article is to analyze how climate change has influenced past evolutionary adaptations and how it is likely to continue impacting future evolution. By doing so, we aim not only to explore historical patterns but also to predict future evolutionary trajectories using machine learning models. This approach allows us to fill a gap in the current literature by addressing the interaction between climate change and adaptive evolution over time, utilizing advanced tools to forecast evolutionary outcomes under various future scenarios.

3 Methodology

In the methodology, it is important to emphasize that in addition to correlating beak dimensions with climatic conditions to verify that climate change is indeed a driving factor in evolution, species classification based on projected climatic conditions and beak dimensions is also conducted. This involves training multiple models with the most recent available data.

3.1 Data Collection

For over four decades, from 1973 to 2012, Peter and Rosemary Grant conducted research on the Galápagos island of Daphne Major, collecting data on Darwin's finches, specifically the species *scandens* and *fortis*, to study Darwin's theory of evolution. During this extensive period, the Grants dedicated themselves to observing, tagging, and measuring Darwin's finches, as well as their environment. Throughout these observations, they documented environmental changes and how these changes benefited certain individuals within the population, exemplifying the process of natural selection in action.

It is worth noting that the climate data was obtained from another research source. The primary focus of the dataset revolves around the dimensions of the finches' beaks. For these birds, the size and shape of their beaks are traits that vary to adapt to changes in their environment. Additionally, it is relevant to mention that the Grants have documented their findings in a book titled "40 Years of Evolution" [17]. It is important to note that this work exclusively utilizes data from the years 1975 and 2012.

3.2 Data Preprocessing

The preprocessing stage involved several steps to prepare the data for analysis: data from the year 2012 was filtered to maintain temporal consistency; feature selection was performed, focusing on temperature, precipitation, beak length, and beak depth; the categorical variable 'species' was converted into numerical values using the 'LabelEncoder' from the 'sklearn' library, specifically for the two species, *Scandes* and *Fortis*; and finally, normalization was applied using 'StandardScaler' to ensure that the features were on a comparable scale.

3.3 Model Selection and Training

Three machine learning models were employed to predict finch species based on climatic conditions and beak dimensions: Logistic Regression, XGBoost, and Random Forest Classifier.

- **Logistic Regression:** A basic Logistic Regression model was implemented to serve as a baseline. This are the key variables in the logistic regression algorithm:

- X : Input features of the dataset.
- y : Labels corresponding to the input features.
- W : Weights applied to the input features.
- b : Bias added to the linear combination of features.
- z : Linear combination of X , W , and b .
- \hat{y} : Model predictions.
- L : Loss between the predictions and the actual labels.
- $\nabla_W L$ and $\nabla_b L$: Gradients of the loss with respect to W and b .
- η : Learning rate, controls the size of updates to W and b .

Algorithm 1 Logistic Regression

Require: Training data (X, y)

Ensure: Optimal weights W^* and bias b^*

1: Initialize weights W and bias b

2: **repeat**

3: Compute the linear combination:

$$z = XW + b \quad (1)$$

4: Apply sigmoid function:

$$\hat{y} = \frac{1}{1 + e^{-z}} \quad (2)$$

5: Compute the loss:

$$L = -\frac{1}{n} \sum (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (3)$$

6: Compute gradients: $\nabla_W L$ and $\nabla_b L$

7: Update weights and bias:

$$W \leftarrow W - \eta \nabla_W L \quad (4)$$

$$b \leftarrow b - \eta \nabla_b L \quad (5)$$

8: **until** convergence

– **XGBoost:** The XGBoost model was configured with hyperparameters tailored to handle multi-class classification, including the objective function ‘multi’ and evaluation metric ‘mlogloss’. The model was trained using 100 boosting rounds with early stopping rounds set to 10. Here are the characteristics of the model:

- X : Input features of the dataset. y : Labels or target values corresponding to the input features.
- $F_T(x)$: The final model generated by XGBoost for making predictions.
- g_t : Gradients of the loss function at round t .
- h_t : Hessians of the loss function at round t .

Algorithm 2 XGBoost

Require: Training data (X, y) **Ensure:** Final model $F_T(x)$

- 1: Initialize model with constant prediction
 - 2: **for** each boosting round t **do**
 - 3: Compute the gradients \mathbf{g}_t and Hessians \mathbf{h}_t
 - 4: Fit a base learner h_t to the gradients and Hessians
 - 5: Update the model: $F_t(x) = F_{t-1}(x) + \eta h_t(x)$
 - 6: **end for**
-

- Base learner h_t : Model fitted to the gradients and Hessians at round t .
- Learning rate η : Controls the contribution of each base model to the final model.

– **Random Forest Classifier:** A Random Forest model was implemented using GridSearchCV to optimize hyperparameters such as the number of estimators, maximum depth, and minimum samples required for splitting and leaf nodes. Cross-validation with 5 folds was used to ensure the robustness of the model. Herea are the parameters of the model:

- (X, y) : Training data. Dataset containing input features
- X and corresponding labels y used for training.
- *param_grid*: Hyperparameter grid. Set of possible values for the model's hyperparameters.
- k : Number of partitions in the cross-validation to assess the model's performance.
- X' : Dataset used to evaluate the trained model's performance.
- \hat{y}' : Class labels predicted by the optimized model \hat{M} for the testing data X' .

Algorithm 3 Random Forest Classification with GridSearchCV

Require: Training data (X, y) , hyperparameter grid *param_grid*, number of folds k **Ensure:** Optimized Random Forest model

- 1: **Input:** Training data (X, y)
 - 2: **Output:** Optimized model \hat{M}
 - 3: Initialize *GridSearchCV* with *RandomForestClassifier* and *param_grid*
 - 4: Perform k -fold cross-validation on (X, y)
 - 5: Select the best hyperparameters based on cross-validation performance
 - 6: Train the *RandomForestClassifier* \hat{M} on the entire training set (X, y) using the best hyperparameters
 - 7: **return** Optimized model \hat{M}
 - 8: **Input:** Testing data X'
 - 9: **Output:** Predicted class labels \hat{y}'
 - 10: Use optimized model \hat{M} to predict class labels for X'
 - 11: **return** Predicted class labels \hat{y}'
-

The performance of each model was evaluated using the following metrics:

- Accuracy Score: The primary metric for evaluating the predictive power of the models.
- Confusion Matrix: To provide a detailed breakdown of the classification performance.

To demonstrate the practical application of the models, future predictions were made based on hypothetical climatic conditions and beak dimensions. For example, predictions were generated for conditions with a temperature of 25 °C, precipitation of 500 mm, beak length of 10 mm, and beak depth of 5 mm.

4 Result and Discussion

The analysis reveals a series of graphical representations that robustly substantiate the phenomenon of evolution, a process notably shaped by climate fluctuations. These visual depictions elucidate the profound influence of climatic dynamics on species attributes, underscoring a continuous evolutionary adaptation to environmental changes.

Furthermore, a predictive study was undertaken to forecast species outcomes by integrating forthcoming beak measurements and climatic variables through three distinct modeling approaches. The findings, outlined in Table 1, succinctly outline the predictive efficacy of each model in anticipating species outcomes amidst diverse environmental conditions. These prognostications offer valuable insights into potential species adaptations and evolutionary trajectories amidst evolving climatic scenarios.

Table 1. Prediction’s results.

Model	Test Accuracy	Train Accuracy	Test Loss	Train Loss
Logistic Regression	0.92	0.97	0.08	0.06
XGBOOST	0.94	0.98	0.09	0.07
Random Forest	0.97	0.99	0.02	0.01

The results demonstrate strong accuracy across all three models. Logistic regression is effectively applied in scenarios with high variable correlation, as evidenced in this study. XGBoost, known for handling complex relationships, also performs admirably in this context. However, the top performer is Random Forest, notably enhanced by the GridSearchCV technique, optimizing hyperparameters.

Furthermore, both test and train accuracies were assessed, with minimal disparity indicating a negligible risk of overfitting. A larger gap between these metrics would typically signal overfitting concerns. An inherent advantage of these models lies in their ability to deliver robust outcomes even with relatively

modest datasets. However, it is evident that one of the limitations of this training is the limited amount of data available, which could also be the reason for the observed degree of overfitting

On another note, when discussing the changes observed in the Galapagos finch beak dimensions between 1975 and the present, a scatter plot was generated for the species *scandens*, as depicted in Fig. 1. This visualization reveals that the beaks were shallower but longer in 1975. Thus, a significant shift is evident, primarily influenced by climatic variations, an assertion that will be substantiated in subsequent graphical analyses.

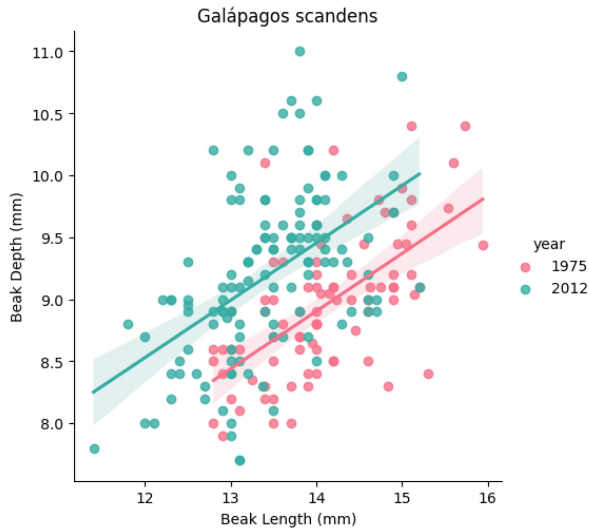


Fig. 1. The plot shows that slope was the same in 1975 and 2012, suggesting that for every millimeter gained in beak length, the birds gained about half a millimeter in depth in both years.

In contrast to the trends observed for the *scandens* species, the *fortis* species exhibited a contrary change, this is in Fig. 2. Between 1975 and 2012, the beaks became shallower but longer, providing another indication of evolutionary progression.

Another analysis conducted was the ECDF (Empirical Cumulative Distribution Function), a valuable tool for visualizing data distributions and comparing the distributions of two or more samples. Figure 3 presents this cumulative probability graph against beak depth for these two species in 1975 and 2012.

The graph for *Fortis* indicates that the probability of finding a *Fortis* finch with a beak depth less than or equal to a certain value was higher in 2012 than in 1975. This implies that in 2012, a larger proportion of *Fortis* finches had deeper beaks compared to 1975. For *Scandens*, the graph shows that the probability of

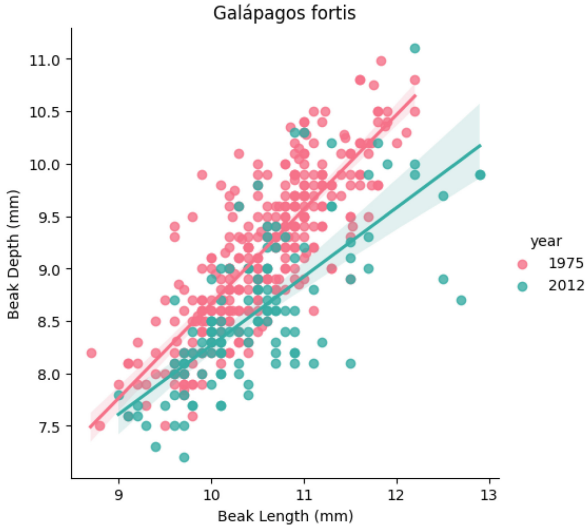


Fig. 2. The comparison indicates that the slope of the 1975 line is steeper than that of the 2012 line, implying a more significant change in beak depth for every 0.5 mm increase in beak length in 1975 compared to 2012.

finding a Scandens finch with a beak depth less than or equal to a certain value was higher in 1975 than in 2012. This implies that in 1975, a larger proportion of Scandens finches had deeper beaks compared to 2012.

The comparison of the ECDF graphs for Fortis and Scandens reveals that the two species have experienced opposite changes in the distribution of beak depth between 1975 and 2012. In Fortis, there has been an increase in the proportion of finches with deeper beaks, whereas in Scandens, there has been a decrease in the proportion of finches with deeper beaks.

Now, when observing the Fig. 4 that compares the cumulative distribution with beak length, it can be seen that for Fortis, it was more likely to find smaller beaks in 1975 and longer beaks in 2012. Conversely, for Scandens, it was more likely to find longer beaks in 1975, but by 2012, these had become shorter.

These changes in the distribution of beak length could have significant implications for the survival and reproduction of the birds. Finches with deeper beaks may be better adapted to consuming certain types of seeds, providing them with an advantage in environments where those seeds are abundant. However, finches with deeper beaks may also be more vulnerable to predators. Various factors must be considered when discussing evolution.

Currently, the discussion has focused on the training results aimed at classifying these species based on their beak characteristics and both recent and future climatic conditions. Graphs have demonstrated noticeable changes in the finches' dimensions over the years. However, the primary goal of this study is to understand why these changes occurred, with climate change being one of the investigated reasons.

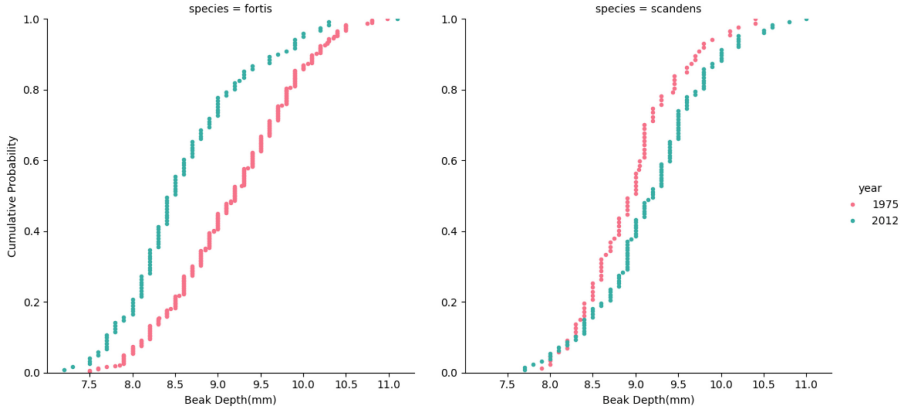


Fig. 3. For *G.fortis* the mean beak depth of 2012 is considerably shorter than 1975. Whereas, for species *G.scandens* the mean beak depth has increased slightly from 1975 to 2012. The variance is slightly higher for *G.scandens* in 2012.

To explore this, a pairplot of all variables in the dataset was created as shown in Fig. 5. This included the beak dimensions of the two species, temperature (in Celsius), and the approximate precipitation levels at the time each species was observed. The objective of generating a pairplot is to visualize the correlations between all variables. At a glance, it is evident that most correlations are highly positive, which aligns with the current understanding that climate change is a reality that has impacted evolution.

It is clear that both precipitation and temperature are highly correlated with the changes in beak dimensions. As climate changes, vegetation changes as well, and species must adapt to survive. Indeed, the finches have changed, as evidenced by their continued existence.

To examine the data numerically and more specifically by species, a confusion matrix was created for both *Fortis* and *Scandens*. In Fig. 6, the confusion matrix for *Fortis* is displayed, where the correlation is even more evident. When a correlation greater than 0.60 is observed, it indicates a positive correlation. This means that as temperature and climate changed, the beak dimensions of *Fortis* also changed.

In the case of the *Scandens* species shown in Fig. 7, the analysis reveals similar patterns to those observed in *Fortis*, albeit with specific differences in evolutionary responses. As previously demonstrated, both species have undergone significant changes in beak morphology, although the directions of these changes are not identical.

It is evident that climate change has also impacted the *Scandens* species, significantly influencing the variables studied. The depth and length of the *Scandens* finches' beaks have shown variations correlated with environmental changes, suggesting an evolutionary adaptation to new climatic conditions.

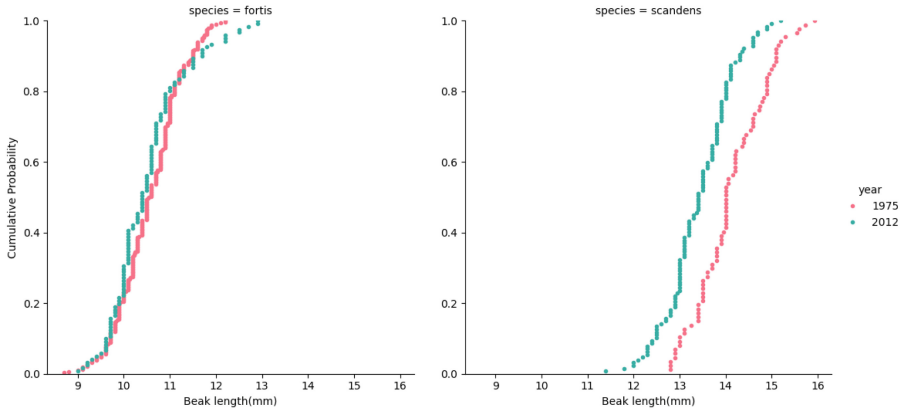


Fig. 4. Observing the preceding graphs, minimal changes in the average beak length of *G. fortis* from 1975 to 2012 are apparent. The 2012 average beak length appears shorter with greater variability. Conversely, the average beak length decreased in 2012 for *G. scandens* species.

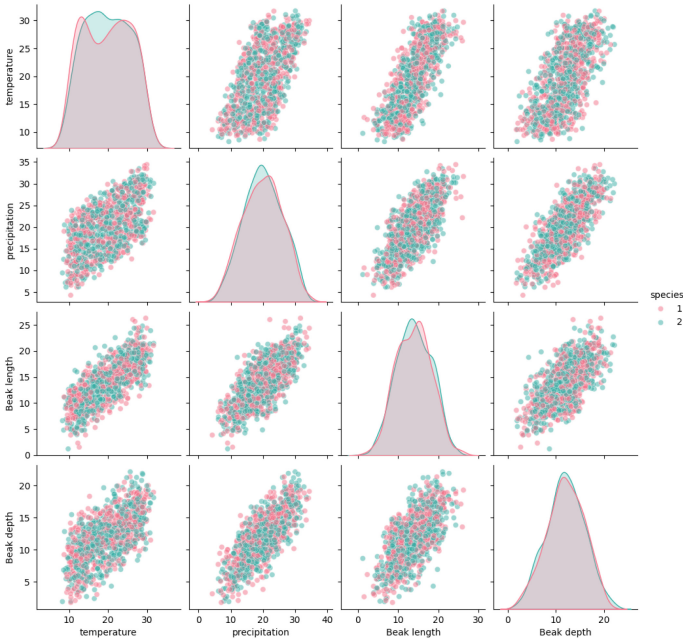


Fig. 5. Here species 1: fortis-species 2: scandens. Relationships between species and climatic conditions in the dataset. This analysis explores how different species interact with various climatic variables. But above all, the variables have a high correlation with each other, which helps us confirm our initial hypothesis: climate change affects evolution.

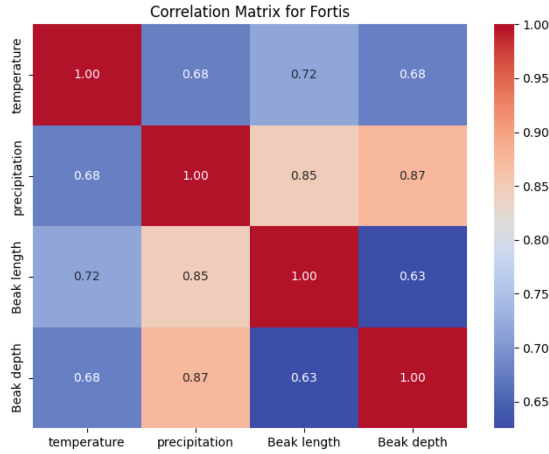


Fig. 6. The results of this correlation matrix suggest that there are a series of complex relationships between the four variables studied. Temperature and precipitation appear to be related, as do peak length and peak depth

These findings underscore the importance of considering the impact of climate change on biodiversity and how different species may respond diversely to the same environmental factors. The relationship between climatic variables and the physical characteristics of Scandens finches indicates a strong connection that warrants further investigation to fully understand the evolutionary dynamics at play.

5 Conclusions

Changes in temperature and other climatic conditions have had a profound impact on the evolution of finches, particularly on the Galápagos Islands, where climate variations have influenced food availability and driven adaptations in beak dimensions. These adaptations have improved the finches' survival and reproductive success across diverse environments. The study used logistic regression, XGBOOST, and Random Forest Classifier to predict finch types based on beak dimensions and updated climatic conditions, providing insights into species adaptation through natural selection. Key points include the significant impact of climate change on finch evolution, the influence of adaptations in beak dimensions due to changes in climate and food availability, and the use of machine learning models for predictions. Future research should focus on applying this methodology to other species and ecosystems, and integrating more advanced machine learning techniques, such as deep learning or ensemble methods, to refine predictions and gain deeper insights into adaptation mechanisms. This approach highlights the remarkable capacity of organisms to respond to changing conditions and underscores the intricate relationship between natural selec-

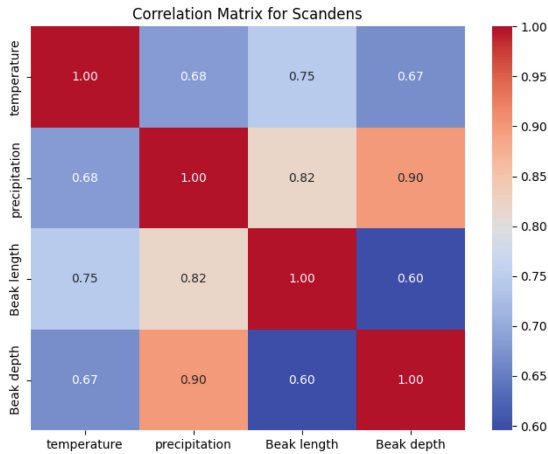


Fig. 7. Temperature also appears to be related to peak length and beak depth, and peak depth also appears to be related to precipitation.

tion and evolutionary change, paving the way for more effective conservation strategies and a comprehensive understanding of adaptation processes.

Additionally, a repository has been created where readers can access the dataset used and the algorithms. This enables them to replicate the results and draw their own conclusions. [GitHub Repository](#)






References

1. Loor, J., Jiménez, A., Aguirre, J., Rodríguez, G., Reyes, I., Vizcaino-Imacaña, P., Morocho-Cayamcela, M.: Artificial intelligence-based detection and prediction of giant African snail (*Lissachatina Fulica*) infestation in the Galápagos islands. In: Proceedings of the 19th International Conference on Software Technologies, ICSOFT, vol. 1, pp. 403–410 (2024)
2. Figueroa, S., Morocho-Cayamcela, M., Pineda, I.: Identifying defective fruits and vegetables with hyper-spectral images: a brief tutorial. In: 2022 Third International Conference on Information Systems and Software Technologies (ICI2ST), pp. 21–27 (2022)
3. Lucero, V., Noboa, S., Morocho-Cayamcela, M.: Plant disease classification and severity estimation: a comparative study of multitask convolutional neural networks and first order optimizers. In: Information Management and Big Data, pp. 313–328 (2022)
4. Camacho, J., Morocho-Cayamcela, M.: Mask R-CNN and YOLOv8 comparison to perform tomato maturity recognition task. In: Information and Communication Technologies, pp. 382–396 (2023)
5. Darwin, C., Murray, J.: On the origin of species by means of natural selection, or. The preservation of favoured races in the struggle for life (1859). <https://doi.org/10.5962/bhl.title.82303>

6. Navalón, G., Marugán-Lobón, J., Bright, J.A., Cooney, C.R., Rayfield, E.J.: The consequences of craniofacial integration for the adaptive radiations of Darwin's finches and Hawaiian honeycreepers. *Nat. Ecol. Evol.* **4**(2), 270–278 (2020). <https://doi.org/10.1038/s41559-019-1092-y>
7. Rivkin, L.R., Johnson, R.A., Chaves, J.A., Johnson, M.T.J.: Urbanization alters interactions between Darwin's finches and *Tribulus cistoides* on the Galápagos Islands. *Ecol. Evol.* **11**(22), 15754–15765 (2021). <https://doi.org/10.1002/ece3.8236>
8. Grant, P.R., Grant, B.R.: Evolution of character displacement in Darwin's finches. *Science* **313**(5784), 224–226 (2006). <https://doi.org/10.1126/science.1128374>
9. Enbody, E.D., Sprehn, C.G., Abzhanov, A., Bi, H., Dobрева, M.P., Osborne, O.G., Rubin, C.J., Grant, P.R., Grant, B.R., Andersson, L.: A multispecies BCO₂ beak color polymorphism in the Darwin's finch radiation. *CB/Curr. Biol.* **31**(24), 5597–5604.e7 (2021). <https://doi.org/10.1016/j.cub.2021.09.085>
10. Carrión, P.L., Raeymaekers, J.A.M., De León, L.F., Chaves, J.A., Sharpe, D.M.T., Huber, S.K., Herrel, A., Vanhooydonck, B., Gotanda, K.M., Koop, J.A.H., Knutie, S.A., Clayton, D.H., Podos, J., Hendry, A.P.: The terroir of the finch: how spatial and temporal variation shapes phenotypic traits in Darwin's finches. *Ecol. Evol.* **12**(10) (2022). <https://doi.org/10.1002/ece3.9399>
11. Beausoleil, M.O., Carrión, P.L., Podos, J., Camacho, C., Rabadán-González, J., Richard, R., Lalla, K., Raeymaekers, J.A.M., Knutie, S.A., De León, L.F., Chaves, J.A., Clayton, D.H., Koop, J.A.H., Sharpe, D.M.T., Gotanda, K.M., Huber, S.K., Barrett, R.D.H., Hendry, A.P.: The fitness landscape of a community of Darwin's finches. *Evolution* **77**(12), 2533–2546 (2023). <https://doi.org/10.1093/evolut/qpad160>
12. Grant, P.R., Grant, B.R.: Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**(5568), 707–711 (2002). <https://doi.org/10.1126/science.1070315>
13. Grant, B.R., Grant, P.R.: Evolution of Darwin's finches caused by a rare climatic event. *Proc. Roy. Soc. Biol. Sci.* **251**(1331), 111–117 (1993). <https://doi.org/10.1098/rspb.1993.0016>
14. Anchundia, D.J., Green, R., Pike, C.L., Gutiérrez, G., Pibaque, P., Chango, R., Sevilla, C., Fessl, B., Tebbich, S.: Habitat restoration to conserve the Little Vermilion Flycatcher *Pyrocephalus nanus* on Santa Cruz Island, Galapagos. *Bird Conserv. Int.* **34** (2024). <https://doi.org/10.1017/s0959270924000091>
15. Reaney, A.M., Bouchenak-Khelladi, Y., Tobias, J.A., Abzhanov, A.: Ecological and morphological determinants of evolutionary diversification in Darwin's finches and their relatives. *Ecol. Evol.* **10**(24), 14020–14032 (2020). <https://doi.org/10.1002/ece3.6994>
16. Al-Mosleh, S., Choi, G.P.T., Abzhanov, A., Mahadevan, L.: Geometry and dynamics link form, function, and evolution of finch beaks. *Proc. Natl. Acad. Sci. U. S. A.* **118**(46) (2021). <https://doi.org/10.1073/pnas.2105957118>
17. Grant, P.R., Grant, B.R.: 40 Years of Evolution: Darwin's Finches on Daphne Major Island. Princeton University Press (2014)



Incident Alert Priority Levels Classification in Command and Control Centre Using Word Embedding Techniques

Marcos Orellana^(✉) , Jonnathan Emmanuel Cubero Lupercio ,
Juan Fernando Lima , Patricio Santiago García-Montero ,
and Jorge Luis Zambrano-Martinez 

Computer Science Research and Development Laboratory (LIDI), Universidad del
Azuay, Cuenca, Ecuador
{marore,flima,jorge.zambrano}@uazuay.edu.ec,
{jcubero,santyg20}@es.uazuay.edu.ec

Abstract. This research analyzes the textual content of emergency calls that arrive at command and control centers to classify them according to their priority. The aim is to respond promptly and make appropriate decisions in situations that require immediate attention. Text mining techniques construct a computational model with preprocessing techniques such as tokenization, case folding, and stop words removal. The calls are then represented using Word Embeddings with the Skip-Gram architecture to obtain word vectorization, and the Clustering algorithm is applied for classification. The results show an improvement in classification accuracy, achieving 95% precision in classification tests using high and low-priority categories and 81% in classification tests using four alert categories. These techniques enhance the syntactic and semantic understanding of emergency calls and reduce the risk of loss of human life.

Keywords: Classification · Clustering · Emergency calls · Machine learning · Text mining · Skip-Gram · Word2Vec

1 Introduction

Healthcare providers use standardized protocols to sort patients based on urgency. These protocols, called triage systems, come in different forms, like the Manchester Triage System (MTS) and the Emergency Severity Index (ESI). Both systems categorize emergencies into five levels, with red being the most critical. Red signifies a life-threatening situation requiring an immediate response. Orange indicates situations needing urgent medical evaluation, while yellow signifies stable but potentially worsening conditions. Green represents non-life-threatening cases that still require attention due to the risk of escalation. Finally, blue signifies emergencies with no severe or life-threatening concerns [3, 18].

Call center operators follow specific protocols when emergencies arise to gather crucial information from the caller. This information helps them determine the severity of the situation, assign the appropriate priority level (red being the most critical), and allocate the necessary resources to address the emergency. This process helps reduce response times [2,4]. In Ecuador, the triage system uses four levels: red for critically ill or injured patients with a chance of survival, orange for those with severe conditions, yellow for non-life-threatening mild illnesses or injuries, and green for minor, temporary issues [15].

Studies present that even with triage protocols, mistakes can happen during emergency call classification [20]. A technique called Word Embeddings can help. It analyzes the conversation between the caller and the operator to understand the meaning and identify keywords. This feature can assist operators in providing the correct response and assigning resources faster.

Word Embedding, a text mining technique, is a step-by-step process that extracts essential terms from a conversation while discarding irrelevant ones. This efficient method is widely used in call centers. It begins by measuring the similarity of words in the conversation, using two main approaches: lexical and semantic. Lexical similarity focuses on the order and similarity of the words themselves, while semantic similarity considers the context in which the words are used, thereby understanding their meaning.

Word Embedding uses vector arrays to represent words and calculate their similarity scores. Tools like *gloVe* or *Word2Vec* can be used for this purpose. Finally, clustering algorithms are used to group similar calls together. These algorithms, often neural networks, analyze the context to predict words and understand the message [1].

Command and Control Centers (C2) are critical for allocating emergency resources. Effective communication between the caller and operator is essential for accurate triage and decision-making. Unfortunately, many C2s face slow response times and inaccurate classifications [15].

The method meticulously analyzes the conversation between the caller and the operator, enabling the categorization of the emergency as high or low priority. This method empowers operators to respond to critical situations swiftly and improve response times. The ultimate aim is to significantly reduce errors in dispatching emergency services (police, medical) and ensure the most severe cases receive the fastest response. This goal should motivate all emergency response professionals.

The present study aims to categorize emergency alerts based on calls transcribed or converted by software to text. The data is classified through machine learning techniques, specifically text mining through Word Embedding. Word Embedding is a technique that represents words in a high-dimensional space, capturing their semantic and syntactic relationships. In this case, the support of technical information helps to complement the development of an investigation to prevent systematization errors.

Finally, this document is organized as follows: Sect. 2 contains the related works, Sect. 3 establishes the methodology of the work, Sect. 4 presents the results, and Sect. 5 contains the conclusions of the work.

2 Related Work

This section describes the work on emergency classification, considering projects related to emergencies and similar proposals. The reference articles are essential for elaborating an adequate distribution of the information received in the C2, used in text mining techniques and clustering algorithms that are appropriate in identifying an emergency. Models or techniques are implemented to eliminate empty words or those that lack value. The techniques used help optimize the alert classification matrices to reduce false emergencies, with efficient results in the prediction of alerts for priority.

Cubranic and Murphy [12] explored classifying errors in medical reports using a Naive Bayes (NB) classifier. Their dataset contained 15,859 entries. The NB classifier attempted to identify classification errors within each report. The data was divided for training and testing purposes. The training data helped the model learn from past examples, while the testing data assessed the model's effectiveness. The researchers experimented with different vocabulary sizes to observe how it affected the results. The initial findings showed the algorithm achieved around 30% precision in assigning correct classifications. Those authors applied preprocessing techniques that abolish words appearing infrequently to improve accuracy. The text processing architecture used during training was Continuous Bag-of-Words (CBOW), and tokenization was applied to separate the text into clean units.

Lee et al. [7] built a model for classifying emergency calls into three situations: first aid, rescue, and disaster. Their model used a decision tree for classification. Before applying the model, they pre-processed the data by removing unimportant words and identifying the most frequent nouns and adjectives in each category. These frequent words helped the model assign a higher probability of belonging to a specific category. This current study differs by using a hierarchical clustering algorithm for classification instead of a decision tree. Hierarchical clustering was chosen because it considers the distance between data points, allowing for more accurate groupings and classifications based on specific characteristics.

Nakata's research [13] uses text mining to categorize accident reports. The method involves dividing the reports into sections such as "start," "engines," "airport," and "control tower". Words were assigned to each category based on their relevance. A clustering algorithm then grouped terms with similar characteristics. Finding a perfect classification system is challenging due to complex data and varying results.

For this reason, this study focused on classifying similar reports using Word Embedding with a CBOW architecture. This technique analyzes word pairs in adjacent sentences, identifying correlations between them throughout the text. The system considers nouns, verbs, adjectives, and proper nouns such as significant words.

Orellana et al. [16] use Word Embeddings with the CBOW architecture to use triage in two categories: high priority and low priority. It analyzes how preprocessing techniques directly correlate with the proposed model's categorization quality and how the frequency of words influences the pattern analysis using the NB classification algorithm. The techniques applied for text processing are similar to research work, which include tokenization, stop word, and lemmatization. A particularly computationally efficient model is Word2vec. The proposed model is the Skip-Gram model, which generally performs better than CBOW. In particular, a Skip-Gram model analyzes the words in a set of sentences (corpus) and attempts to predict which words are neighbors.

Trujillo et al. [19] focus on C2 such as Integrated Security Service (ECU 911). These centers rely on Computer-Aided Dispatch (CAD) systems to manage emergencies. While these systems help register incidents and allocate resources, they can suffer from information loss. For this reason, this study proposes a new system to address this issue through Natural Language Processing (NLP) techniques and algorithms to automate call registration, such as Automatic Speech Recognition (ASR) [14], Named-Entity Recognition (NER), Term Frequency and Inverse Document Frequency (TF-IDF) (This measure demonstrates whether a word becomes relevant to a document in a collection of terms), and Support Vector Machine (SVM).

The information in the previous articles demonstrates a connection and similarity with the purpose of this work. Some models focus on search patterns to classify priorities or alerts, and the categorizations used need to be completed triages. In other cases, they are not oriented toward alerts, and the data is aimed at studying documents or medical reports that contain formal language. Therefore, our research aims to develop a machine learning-based system for categorizing emergency alerts. The process involves converting audio emergency calls into textual data, which is then subjected to text mining using Word Embedding. This technique projects words into a high-dimensional space to capture semantic and syntactic relationships, enabling the classification of emergency alerts. By leveraging technical information, the study seeks to enhance the accuracy of the classification model, mitigating potential systematization errors.

3 Materials and Methods

A methodology that describes the classification process of systematic techniques has been used to obtain results. This methodology predicts the priority of emergency calls through the Word Embedding method, which transforms natural language into proximity vectors, and the clustering algorithm is used to group objects according to their similarity [11]. For this analysis, the Word Embedding

techniques were applied with the Word2vec tool, which distributes an automatic association of information under test. The input is composed of the dataset for the NLP, transforming each word into vectors. The Word Embedding technique applies the Skip-Gram architecture, a model that analyzes the corpus based on fragments, which takes “n” previous fragments and “n” subsequent fragments, evaluates them, and creates knowledge patterns [8].

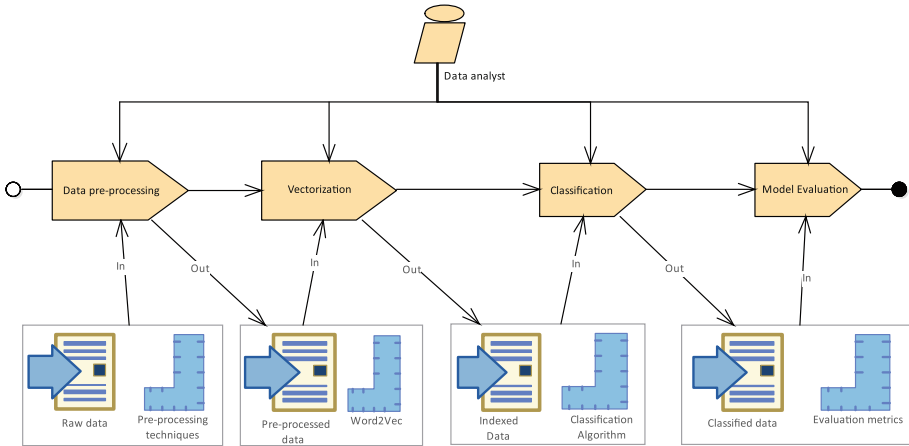


Fig. 1. Systematic techniques process to obtain results.

Figure 1 presents the research methodology developed through the Systems Process Engineering Metamodel 2.0 (SPEM 2.0). The processes demonstrate the convergence of the starting method and describe the inputs and outputs in each research development process. The methodology was divided into the following processes: (i) Data Preprocessing, (ii) Vectorization, (iii) Classification, and (iv) Model Evaluation.

3.1 Data Preprocessing

The input artifacts in this activity were the original dataset and the preprocessing techniques. The original dataset represents the corpus to train the model, highlighting its size, characteristics, and origin. Preprocessing techniques are used to obtain helpful information and enrich the content. This investigation was based on emergency calls in Spanish from the ECU 911 and the transcriptions made by Computer Science Research and Development Laboratory (LIDI).

This corpus of transcribed calls consists of approximately one thousand emergency calls. The corpus was classified into two categories: high alert (in this alert, there is only a red priority) and low alert (this alert includes yellow, orange, and green priority). For the second case study, 512 were classified as red priority, 354 transcripts as orange priority, 36 as yellow priority, and 96 as green priority, as

presented in Table 1. Furthermore, the Test column specifies the development or training applied for the respective categories.

Table 1. Representation of triage of two and four categories.

Case	Category	Priority	No. calls	Test
1st	High alert	Red	512	Test 1 - Test 2
	Low alert	Orange, Yellow, Green	486	Test 1 - Test 2
2nd	Red alert	Red	512	Test 3 - Test 4
	Orange alert	Orange	354	Test 3 - Test 4
	Yellow alert	Yellow	36	Test 3 - Test 4
	Green alert	Green	96	Test 3 - Test 4

Raw Data The dataset used in this process is raw data, known as unprocessed primary data, as it is initially taken from the source. This dataset has no transformation; these data are units of information fragments. This process begins by extracting the text of the Spanish calls from the ECU 911.

Preprocessing Techniques In the transcription of emergency calls from ECU 911, the text contains frequent errors [14]. Generally, these texts have colloquial language or idioms with a non-formalized structure or are erroneous and informal and also contain acronyms, symbols, and grammatical errors.

These problems are resolved using preprocessing techniques to improve the quality of the information and eliminate noise and frequent grammatical errors by using calls from colloquial conversations as a corpus, specifically calls from the alert [10]. Therefore, it is advisable to use preprocessing techniques to eliminate specific problems, such as linguistic mismatch, where the words in a query do not match, and thus obtain transcripts with the necessary information to minimize possible errors in the model. The present study applies tokenization, case folding, stop words, and lemmatization.

- **Case Folding:** Text calls contain proper names, places, and addresses, usually in lowercase and uppercase. The letters of the dataset are represented by ASCII code. Those codes interpret letters in both lower and upper case differently. Case folding converts upper case letters to lower case letters in the dataset. It is recommended that this technique be used in all texts.
- **Tokenization:** Once the structure is obtained, the text conversations are separated into tokens; these tokens are considered entries or strings where punctuation marks are not considered tokens.
- **Stop Words:** Prepositions, articles, pronouns, and other words are frequently used in emergency calls. These linguistic components do not provide any helpful information, so if repeated excessively, they are eliminated or replaced to give the text coherence and naturalness.

- **Stemming and Lemmatization:** This process reduces inflection, such as derived word forms, to a joint base. Usually, the Stemming technique refers to the heuristic process of cutting off the ends of a word to reduce it.

3.2 Vectorization

The text data are clean or already processed at this step, and the words are placed in a vector matrix. The matrix is formed of rows, which are represented by converted words. The row is the vector divided into boxes, where each box represents a numerical value of a word in the text. The entire data set is now converted into the proper representation for Word2vec.

Word2vec The Word Embeddings technique is reflected in the application of the Word2vec tool, and there are two learning algorithms for this technique: CBOW and Skip-Gram [6]. Both methods use the concept of a neural network that maps words to target variables, which are also words. In these techniques, *weights* are used as word vector representations. CBOW tries to predict a word based on its neighbors, while Skip-Gram tries to predict the neighbors of a word. It is also important to note that the weight of the matrix between the input and the projection is shared for all positions of the abstracted vocabulary [9]. This research uses a Skip-Gram model as the base architecture, which takes words before and after each word in the conversation to find similarities by creating the context and providing an array of decimal numbers representing each similarity. In addition, the dimensionality of each variety is presented; in that case, the value increases, and the Embeddings quality of each word is determined. The quality gain decreases after reaching an inflection point.

3.3 Classification

Different algorithms can be used to classify texts into predefined categories. Three types of classification must be considered to select one of these algorithms: unsupervised, supervised, and semi-supervised methods [5]. In this case, a method was used to perform a supervised analysis.

Index Data and Classification Algorithm Once the vectorization stage is completed, the data is indexed for training. These data are input to the supervised training stage, and relevant features are extracted. At this stage, clustering is responsible for grouping texts according to the similarity that exists between them. The cluster creates groups from the transcribed texts. It is considered that automatic classification orders texts according to specific pre-established categories. The dataset in this use case, the C2, keeps emergency calls in documents, and by text mining algorithms, it can group these documents and retrieve helpful information.

This study focuses on categorizing alerts based on the labels in the dataset. For this classification type, the number of clusters is configured depending on the

number of classifications in the dataset, using the cluster-based algorithm and the TF-IDF technique are considered. The initial test was configured with two clusters when two categories were classified (high priority and low priority), and four clusters (red, orange, yellow, and green) were defined for the second test. Once each word was classified into clusters, a centroid was obtained, but it cannot be identified which cluster belongs to which category; it only has groupings of similar texts.

That problem can be solved with a filter applied in the dataset for each category, and through the TF-IDF technique, the words with the highest index belonging to a category in the dataset were found. Thus, it was compared with the words assigned to each cluster, searching for the more significant similarity between the words with the highest membership index. As a result, the cluster that belongs to each category was found.

Figure 2 presents the groups considered in this research distributed in the same two-dimensional space. This aspect was considered the group of characteristics in Test 1 and Test 2. In Word2Vec, there are 84 dimensions, which are reduced to two to be represented graphically. Test 1 represents the data of two categories, which are high and low alert. The dataset was available with many attributes and instances since it is necessary to reduce the data to more minor variables or cases that are the alerts, losing the lowest possible quality of information; otherwise, the runtime can be high, as seen in Fig. 2. In Test 1, there are two different groups, 0 and 1. These are elements that are grouped when similarities are found among them. The same happens with the value 1, which looks for similarities and forms a specific group representing low and high alerts. This behavior occurs also in Test 2 of four categories. They are grouped depending on the elements' characteristics and form groups represented in 2D.

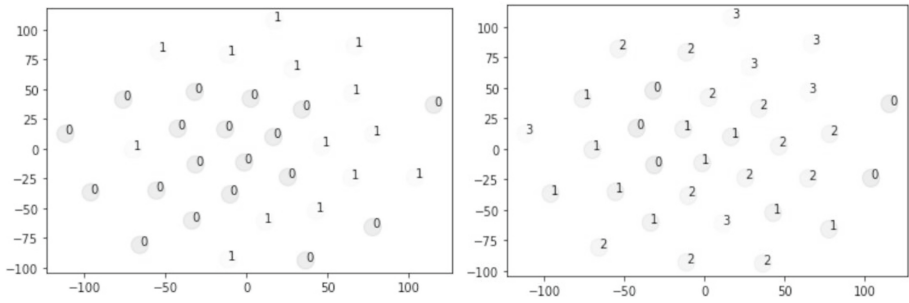


Fig. 2. Enhanced classification performance with clustering and dimensionality reduction

3.4 Model Evaluation

Evaluating a classification model involves several analyses to determine the validity of the ECU 911 emergency alert model. The evaluation reduces sending an

erroneous resource when an ambulance is needed, saving human lives. Therefore, once the data categorization stage has passed, the results must be evaluated in the 2×2 confusion matrix for the first classification of two categories; below, the matrix is described of confusion.

Confusion Matrix In the pattern recognition process, the confusion matrix is applied to evaluate performance regarding successes and errors in the new data classification phase. Table 2 describes the confusion matrix of the two-category classification system with the highest weight in recall. Table 3 describes the same matrix; the difference is the precision weight. These expressions are defined for the confusion matrix: (i) True Negative (TN), (ii) True Positive (TP), (iii) False Negative (FN), and (iv) False Positive (FP).

Table 2. Values with the greatest weight in recall.

Classification	Positive prediction	Negative prediction	Precision
High alert	484	465	51%
Low alert	28	21	42.85%
Recall	95%	4.32%	–
F1-score	66%	7%	–

Evaluation Metrics Evaluating a machine learning model is an essential part of any project. The model can provide satisfactory results when evaluated with one metric. Most of the time, classification accuracy is used to measure the models’ performance, but more is needed to judge the model truly. The metrics that were used in this study were:

- **Recall.** The metric that reports the number of TP that the machine learning model can use to identify the type of alerts (red, green, yellow, or orange). Thus, there are 998 alerts, and predicting which ones are positive is necessary. Despite not being a huge sample, it is robust for machine learning models. The responsibility for handling this sensitive information is on the part of the LIDI with different confidentiality agreements to manipulate the sensitive data of the transcribed call information. This involves having legal agreements for this dataset and its exclusive use for LIDI research. According to this metric, 92.96% of the positive cases are captured, which indicates that 4.5 out of 5 cases are correct.
- **Precision.** This metric calculates the percentage of cases in which the model has been successful. The disadvantage of this precision approach is that it is arguably misleading, and the model cannot be trained correctly. Precision predicts 92.96% of cases are correct, but applying a recall metric is recommended when the data is unbalanced.
- **F1-score.** The harmonic mean of precision and recall is calculated. The result of this metric is 78.80% of the calculation rate.

Table 3. Values with the greatest weight in precision.

Classification	Positive prediction	Negative prediction	Precision
High alert	476	36	93%
Low alert	220	266	54.73%
Recall	68%	88%	–
F1-score	79%	68%	–

4 Results and Discussions

The results obtained after the set of tests demonstrate an improvement in the recall performance parameter, which results in 81% for the first and second tests and 86% for the third test, concerning the study described in [15].

This study uses a dataset of 998 call records from ECU 911 and implements the vectorization technique using the Word2vec tool combined with the Skip-Gram architecture. The evaluation metric obtained is 95% in Test 1 in the first training (Table 2) and 68% in retraining (Table 3).

These results were the evaluation for the two categories and 88.44% for Test 3 of the four categories of the first training (Table 4); for retraining, a 24.81% evaluation rate was obtained (Table 5).

Table 4. Four alert confusion matrix, first training.

	Yellow	Red	Orange	Green	Support	Precision
Yellow	30	1	1	1	36	90.90%
Red	1	444	56	11	512	86.71%
Orange	10	45	245	54	354	70%
Green	2	12	22	60	96	62.5%
Recall	69.76%	88.44%	75.61%	47.61%		
F1-score	79%	88%	72%	54%		

This study used preprocessing techniques to eliminate less frequent words, symbols, and empty spaces, passing them as clean data, indexed data, and vectorizing with the Word2vec tool.

To classify the model into two categories, a clustering algorithm was used with the TF-IDF technique along different ranges of words and configurations, thus achieving the best possible recall.

Likewise, the same method was used to classify the four categories (Test 2), but changing the number of clusters was implemented in Test 1. The dataset was divided into two categories (high priority and low priority), and for Test 2, four categories were defined (red, orange, yellow, and green). Test 1 and Test 2

Table 5. Four alert confusion matrix, retraining.

	Yellow	Red	Orange	Green	Support	Precision (%)
Yellow	416	81	10	5	512	81.25%
Red	1	33	1	1	36	91.66
Orange	5	6	313	30	354	88.41
Green	10	13	6	67	96	69.79
Recall (%)	96.29	24.81	94.84	65.05		
F1-score (%)	88	39	92	67		

have their respective retraining, adjusting the rate in the evaluation metrics, as shown in Table 5. Therefore, higher performance is evaluated in Table 5, where the values are higher than in the first training.

The study compares two models (T1 and T2) for classifying emergency alerts. Figure 3 shows that T1 performs slightly better in recall (identifying true emergencies) but worse in F1-score (overall accuracy) than T2. This behavior is because T2 prioritizes precision (avoiding false alarms), which reduces recall. Both models use precision, recall, and F1-score to evaluate alert classification. High recall is crucial for red alerts (life-threatening emergencies). This evaluation method helps identify errors in model configuration and optimize its performance.

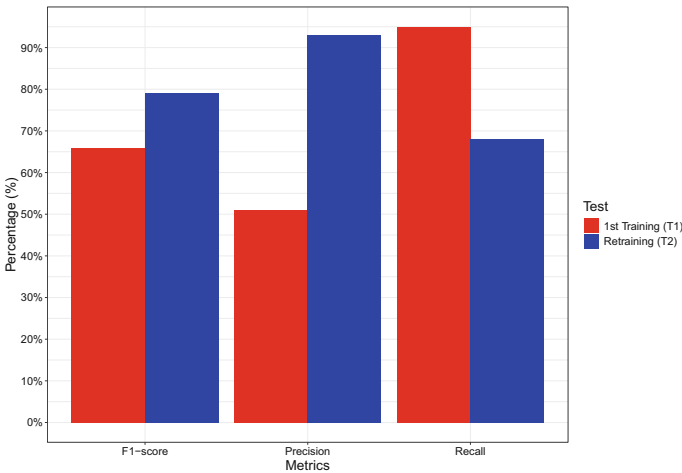


Fig. 3. Evaluation metrics for the two categories classification (High - Low Priority) of the first training (T1) and retraining (T2)

In the current work, tests were performed with a classification of four categories, which resulted in 88.44% of the evaluation rate (Fig. 4). Concerning

Orellana et al. [15], the recall rate as an evaluation measure was equal to 86%. T1 was tested with two categories whose prediction was 95% in the evaluation indicators (Fig. 3). However, similar results were not reached due to the difference of 107,609 trained documents between the two studies.

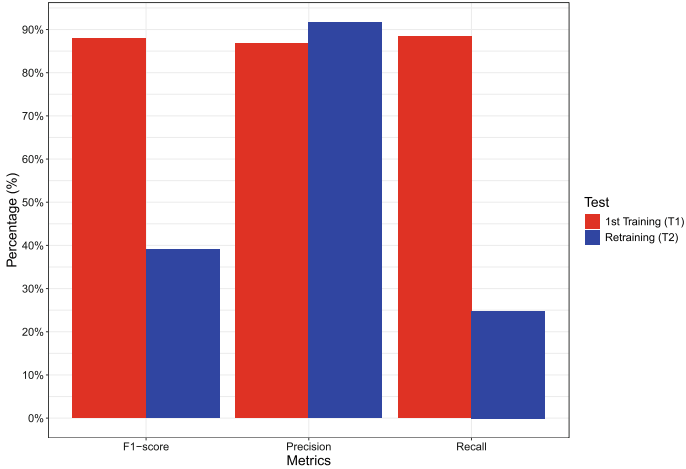


Fig. 4. Evaluation metrics for the four categories classification (red, orange, yellow, and green) of the first training (T1) and retraining (T2)

The benefit of this approach is that it supports emergency response by improving response times, resource allocation, and communication between callers and operators. Tables 6 and 7 show how the model prioritizes recall over precision. Training the model with more weight on recall reduces the ability to identify true emergencies for the most critical (red alerts) but increases precision for less critical ones. This trade-off is acceptable for this study since recall is the main focus.

Similar to Murphy and Cubranic [12], where accuracy decreased as the size of the test dataset increased, the metrics found behave similarly. Dividing the dataset into two categories (T1) increased the metrics in T2 (four categories) because some high-priority cases were shifted to lower urgency categories in T2 (orange, yellow, and green). The study demonstrates that the proposed model focusing on recall can effectively classify emergency alerts.

Table 6. The dataset of two categories with the highest evaluation rate for recall.

Classification	Precision (%)	Recall (%)	F1-score (%)	Number of cases
High	51	95	66	512
Low	43	4	7	486

Table 7. Dataset of four categories for T3 values.

Classification	Precision (%)	Recall (%)	F1-score (%)	Number of cases
Yellow	91	70	79	512
Red	87	88	88	36
Orange	69	76	72	354
Green	63	48	57	96

Two previous studies explored training models for emergency response systems. Roldan [17] compared self-training and co-training approaches, achieving both low false negative rates (1.17 and 1.18%). In contrast, this study prioritizes recall (correctly identifying true emergencies) over precision (avoiding false positives). Although the recall for red alerts in T3 is lower (24.81%), the precision is higher (88%) (Fig. 3 and Table 7).

Orellana et al. [16] investigated reducing the dimensionality of infrequent words in their model for high and low-priority calls. Their recall reached 91% with 75% precision after dimensionality reduction. The current approach utilizes clustering to classify alerts after preprocessing the text to minimize errors. The results using Word Embedding and other techniques show a success rate of 88% for red alerts, followed by orange (76%), yellow (70%), and green (48%) (Table 7). When focusing on recall for high and low priorities (Table 6), the model performs better than T2 (1% improvement) with 95% and 94% recall, respectively. This value also surpasses the 91% recall Orellana et al. [16] achieved.

The study emphasizes recall because we prioritize identifying true emergencies in emergency calls. However, F1-scores (a combined measure of precision and recall) of 95% and 94% for high and low priorities in T1 (Table 6) were still good results. In T2, F1-score-based classification resulted in lower performance for yellow, orange, and green alerts than for red ones. This behavior is why T3 balances the results across categories, except for green alerts (48%). T1 also achieved F1 scores above 50% for high and low-priority alerts.

5 Conclusion

This study evaluates the application of the Word Embedding technique to improve the classification of emergency alerts in a C2. Two data sets were used, one with two categories of alerts and another with four categories. Different NLP techniques were applied, and the model’s performance was evaluated using metrics such as recall and F1-score. Word Embeddings significantly improves classification accuracy, especially in reducing false positives for red alerts. A recall of 95% was obtained for two categories and close to 95% for four categories. It was found that a more robust data set could further improve the recall above 88%. The F1-score can be enhanced (reaching 48% or more) through further preprocessing and enrichment of the data set, especially for categories with less data.

Evaluation metrics can never be perfect due to the complexity of natural language and the large amount of data that must be processed. It is essential to update the NLP techniques used to handle noise in the data better, as well as idioms, colloquial language, and various expressions in Spanish of different regions. Misclassification of alerts can have negative consequences, such as misallocation of resources. Reducing triage errors is crucial to ensuring rapid and efficient response to emergencies. Word Embedding are an effective technique for improving emergency alert classification. The implementation of this technology in C2 can save lives and optimize the use of resources. Authorities must invest in training and implementing new technologies to improve emergency management.

For future work, the approach could propose Word Embedding such as *FastText* and *Glove* which serve as a basis for representing words and classification. Also, this approach can be compared and improved in accuracy with other modern techniques using Transformer architectures. Likewise, use text analysis and NLP in RStudio using the *tidytext* and *dplyr* libraries, an investigation that improves the models presented in this work.

Acknowledgments. This work was supported by the vice rectorate of Research at Universidad del Azuay. Therefore, we thank them for their financial and academic support and the entire Computer Science Research & Development Laboratory (LIDI) staff.

References

1. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al.: Semeval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 252–263 (2015)
2. Chong, H., Gan, K.: Development of automated triage system for emergency medical service. In: 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES), pp. 642–645. IEEE (2016)
3. Hinson, J.S., Martinez, D.A., Schmitz, P.S., Toerper, M., Radu, D., Scheulen, J., Stewart de Ramirez, S.A., Levin, S.: Accuracy of emergency department triage using the emergency severity index and independent predictors of under-triage and over-triage in Brazil: a retrospective cohort analysis. *Int. J. Emerg. Med.* **11**, 1–10 (2018)
4. Inoue, S., Sonoda, A., Yasuura, H.: Triage with RFID tags for massive incidents. In: RFID Handbook, pp. 329–349. CRC Press (2017)
5. Korde, V., Mahender, C.N.: Text classification and classifiers: a survey. *Int. J. Artif. Intell. Appl.* **3**(2), 85 (2012)
6. Kuang, S., Davison, B.D.: Class-specific word embedding through linear compositionality. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 390–397. IEEE (2018)
7. Lee, K., Kim, J.K., Park, M.W., Kim, L., Hsiao, K.F.: A situation-based dialogue classification model for emergency calls. In: 2017 International Conference on Platform Technology and Service (PlatCon), pp. 1–4. IEEE (2017)

8. Meyer, D.: How exactly does word2vec work. Uoregon.Edu, Brocade.Com, pp. 1–18 (2016)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
10. Møller, T.P., Kjærulff, T.M., Viereck, S., Østergaard, D., Folke, F., Ersbøll, A.K., Lippert, F.K.: The difficult medical emergency call: a register-based study of predictors and outcomes. *Scand. J. Trauma Resuscitation Emerg. Med.* **25**, 1–9 (2017)
11. Morissette, L., Chartier, S.: The k-means clustering technique: General considerations and implementation in Mathematica. *Tutor. Quant. Methods Psychol.* **9**(1), 15–24 (2013)
12. Murphy, G., Cubranic, D.: Automatic bug triage using text categorization. In: *Proceedings of the Sixteenth International Conference on Software Engineering & Knowledge Engineering*. pp. 1–6. Citeseer (2004)
13. Nakata, T.: Text-mining on incident reports to find knowledge on industrial safety. In: *2017 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–5. IEEE (2017)
14. Orellana, M., Jiménez Sarango, Á.A., Zambrano Martínez, J.L.: Improving audio of emergency calls in Spanish performed to the ECU 911 through filters for ASR technology. In: *X Jornadas de Cloud Computing, Big Data & Emerging Topics, Vol. 2022. La Plata* (2022)
15. Orellana, M., Trujillo, A., Acosta, M.I.: A methodology to predict emergency call high-priority: case study ECU-911. In: *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*. pp. 243–247. IEEE (2020)
16. Orellana, M., Trujillo, A., Acosta, M.I.: Improving classification models using the frequency of terms and a percentage relation approach between classes for emergency calls. In: *Future of Information and Communication Conference*, pp. 216–226. Springer (2021)
17. Roldán Bocanegra, I.: *Aprendizaje Semisupervisado*. Universidad de Sevilla, Sevilla (2020)
18. Storm-Versloot, M.N., Ubbink, D.T., Kappelhof, J., Luitse, J.S.: Comparison of an informally structured triage system, the emergency severity index, and the Manchester triage system to distinguish patient priority in the emergency department. *Acad. Emerg. Med.* **18**(8), 822–829 (2011)
19. Trujillo, A., Orellana, M., Acosta, M.I.: Design of emergency call record support system applying natural language processing techniques. In: *Conference on Information Technologies and Communication of Ecuador*, pp. 53–65. Springer (2019)
20. Zhang, J., Zhang, M., Ren, F., Yin, W., Prior, A., Vilella, C., Chan, C.Y.: Enable automated emergency responses through an agent-based computer-aided dispatch system. In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1844–1846 (2018)



Physics Informed Neural Networks and Gaussian Processes-Hamiltonian Monte Carlo to Solve Ordinary Differential Equations

Robert Chachalo¹ , Jaime Astudillo¹ , Saba Infante¹ ,
and Israel Pineda² 

¹ Universidad Yachay Tech, Urcuqui, Ecuador
{roberth.chachalo, jaime.astudillo, sinfante}@yachaytech.edu.ec

² Universidad San Francisco de Quito, Quito, Ecuador
ipineda@usfq.edu.ec

<https://www.yachaytech.edu.ec>

<https://www.usfq.edu.ec/es>

Abstract. Non-linear systems of differential equations are vital in fields like biology, finance, ecology, and engineering for modeling dynamic systems. This paper explores two advanced function approximation techniques Physics Informed Neural Networks (PINNs) and Gaussian Processes (GPs) combined with Hamiltonian Monte Carlo (HMC) for solving Ordinary Differential Equations (ODEs) that represent complex physical phenomena. The proposed approach integrates PINNs and GP-HMC, demonstrated through two synthetic models (Lotka Volterra and Fitzhugh Nagumo) and a real dataset (COVID-19 SIR model). The results show that the methodology effectively estimates parameters with low Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). For example, in the Lotka-Volterra model, GP-HMC achieved an RMSE of 0.044 and MAE of 0.041 for one state variable, while PINNs yielded an RMSE of 0.106 and MAE of 0.081. These results highlight the robustness of the methodology in accurately reconstructing system states across varying levels of variability.

Keywords: Physics-Informed Neural Networks · Gaussian Processes · Hamiltonian Monte Carlo · Ordinary Differential Equations · Bayesian Inference · Uncertainty Quantification

1 Introduction

There is an immense interest in making inferences and prediction of complex real-world processes, problems that arise in science, engineering and industry. The ordinary differential equations (ODEs) are used to study complex dynamic phenomena or dynamical systems, by example are of interest in fields such as biology, finance, ecology and biochemistry. Modern deep learning methods are incredibly powerful tools for tackling a host of complex problems.

Deep learning has led to a revolution in machine learning, providing solutions to tackle problems that were traditionally difficult to solve. In recent years, the integration of machine learning techniques with classical numerical methods has revolutionized the field of computational science, particularly in solving differential equations that model complex physical phenomena. Among these innovative approaches, Physics-Informed Neural Networks (PINNs) [1–5] and Gaussian Processes (GPs) [6, 7] have emerged as powerful tools for addressing challenges inherent in the numerical solutions of Ordinary Differential Equations (ODEs) and Partial Differential Equations (PDEs). PINNs leverage deep learning architectures to encode underlying physical laws that govern the behavior of systems described by ODEs [1, 14–17]. These networks not only predict outcomes but are also trained to obey the laws encapsulated in the differential equations, thus ensuring that predictions remain physically plausible. This methodology aligns closely with the principles of scientific computing where accuracy and adherence to physical laws are paramount. The research detailed in [17, 22] exemplifies the deployment of PINNs to model second-order ODEs with sharp gradients, a common challenge in engineering applications involving rapid changes in boundary conditions or material properties.

On the other hand, Gaussian Processes (GPs) [6] are used to model functional data, because they are flexible, robust to outliers, and provide a calibrated uncertainty estimate. Deep Gaussian Processes (DGP) are a generalization of a multilayer neural network seen as a GP in the limit [8, 26]. GPs offer a probabilistic approach to solving ODEs, providing not just predictions but also quantifying the uncertainty inherent in those predictions [7, 24]. This is particularly valuable in scenarios where data is sparse or noisy, as is often the case in real-world measurements. GPs have been effectively applied to model complex systems where the underlying dynamics are not fully understood, making them invaluable for exploratory analysis and for situations where model flexibility is crucial. We combine this probabilistic technique that allows automatic Bayesian inference, with Gradient-based algorithms for Markov chain Monte Carlo (MCMC) sampling, known as Hamiltonian Monte Carlo (HMC) [9]. HMC is a sampling algorithm that was originally developed for molecular dynamics [10–12]. It is now commonly used for sampling problems where the gradients of the posterior probability distribution $P(\theta|y)$ with respect to the model parameters θ are easy to compute, allow inference on increasingly complex models but requires gradient information that is often not trivial to calculate.

The synergy between PINNs and GPs in solving ODEs lies in their complementary strengths—PINNs bring structure and physics-based constraints, while GPs introduce flexibility and uncertainty quantification. The contribution of this article consists of integrating two methodologies based on deep neural network (DNN) techniques and GP to approximate solutions to ODE. Showing the power of these two tools that provides robust solutions, improves prediction accuracy, provides clear analysis of uncertainty, and addresses complex problems.

The rest of the article is as follows: Sect. 2 describes the problem; Sect. 3 defines the methodology; Sect. 4 shows the results; and Sect. 5 shows the Discussion and conclusion.

2 Problem Description

This study aims to develop a robust methodology for solving non-linear ODEs that model complex physical phenomena by integrating Physics-Informed Neural Networks (PINNs) with Gaussian Processes (GPs) and Hamiltonian Monte Carlo (HMC). The approach is tested on synthetic models (Lotka-Volterra, Fitzhugh-Nagumo) and a real dataset (COVID-19 SIR model), demonstrating its effectiveness in accurate parameter estimation and uncertainty quantification.

We study the problem of finding a real-valued curve $x : T \rightarrow \mathbb{R}$ over an interval $T = [t_0, T]$ such that

$$\dot{x}(t) = \frac{dx(t)}{dt} = f(x(t), t, \theta), \quad x(t_0) = x_0 \tag{1}$$

where θ is a vector of unknown parameters, with f Lipschitz continuous with constant C in the second argument and sufficiently many times differentiable in its second argument and $x(t_0) = x_0$ the initial value. A problem that appears frequently, in many areas, consists of: given $x(t_0)$, calculate $x(t_1)$, i.e. an initial value problem, whose formal solution is:

$$x(t) = x(t_0) + \int_{t_0}^t f(x(t), t, \theta) dt \tag{2}$$

In most problems the integral given in Eq. (2) cannot be solved analytically, so one has to resort to Numerical Methods, Markov chain Monte Carlo (MCMC), or Sequential Monte Carlo (SMC). Deterministic numerical methods for the integration of this Eq. (2) on time interval $[0, T]$ will produce an approximation to the equation on a mesh of points $\{t_k = kh\}_{k=0}^K$, with $Kh = T$

$$x(t) = ODEsolve(f(x(t), t, \theta), t_0, t, \theta). \tag{3}$$

In equation (1) there are three problems to solve: find solution states $x(t_1), \dots, x(t_k)$, estimate the parameters θ and predicting the value of an unknown function $\hat{y} = \hat{f}(x(t), t, \theta)$.

The methodology enhances the accuracy and reliability of solving non-linear ODEs, which is crucial in fields like biology, engineering, and finance. By combining PINNs’ physics-based constraints with GPs’ probabilistic modeling, the study provides a powerful tool for addressing complex problems where traditional numerical methods fall short.

3 Methodology

This work proposes using a methodology based on a regression version of neural network to approximate the function f in (2). Machine learning and statistics provide other than regression models. For example, neural networks and Gaussian processes incorporating the knowledge of the equation. Neural networks

(NN) are widely used to solve problems in a variety of domains including computer vision, game theory, classification of an image, speech recognition, natural language processing, as well as computational advertisement, etc. [18, 19]. In particular, we will use a methodology based on physics-informed neural networks (PINN), which is a scientific machine learning technique used to solve problems involving ordinary differential equations (ODEs) and partial differential equations (PDEs) [1, 2, 15–17, 20, 21], and [22]. PINNs approximate PDE solutions by training a neural network to minimize a loss function.

The use of machine learning approaches in the field of scientific computing including differential equations is relatively recent. The abstract goal of approach is to find a function $f : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $f(\cdot, \theta)$ accurately predicts the result of an observed phenomenon. This can be formalized as the initial value problem

$$\frac{dx(t)}{dt} = \hat{f}(x(t), t, \omega), \quad x(t_0) = x_0 \tag{4}$$

from a time series $\{x(t_k)\}_{k=1}^n$, with $x \in \mathbb{R}^n$, non-uniformly sampled, from the observation of the unknown system (2), where \hat{f} is a function parameterized by a neural network (NN). The function \hat{f} is parameterized by the weight vector $\omega \in \mathbb{R}^p$ that is trained using examples. We can chosen the space of functions, of which \hat{f} is an element, and defined a norm in that space, [23] to set a optimization problem:

$$\omega := \arg \min_{\omega} [\mathcal{L}(\epsilon, \omega)] \tag{5}$$

where

$$\mathcal{L}(\epsilon, \omega) = \frac{1}{2} \left(\sum_{k=1}^{n-1} \epsilon_k^2 + \lambda \|\hat{f}(x(t), \omega)\| \right) \tag{6}$$

and

$$\epsilon_i = x(t_{k+1}) - x(t_k) + \Delta t_k \hat{f}(x(t_k), \omega) \quad \Delta t_k = t_{k+1} - t_k \tag{7}$$

Let us now introduce an ODE in the following residual form, [20]

$$\mathcal{F} \left(t, x, \frac{dx}{dt}, \frac{d^2x}{dt^2}, \dots \right) = 0, \quad t \in [t_0, T] \tag{8}$$

The NN is used to evaluate the residual of the ODE equation via x_θ and corresponding derivatives. Let $x(t)$ and its approximation $\hat{x}_\theta(t)$. The input first layer of the neural network is supplied with N_d data values at different times $t_i \in [t_0, T]$ corresponding to $x_d(t_i) = x_d^{(i)}$ data with $i = 1, \dots, N_d$. The initial condition corresponds to the first point, $x_d(t_0) = x_d^{(1)}$ data. For simplicity, we assume a uniform distribution of the N_d points within a sub-interval. The two partial loss functions \mathcal{L}_d and $\mathcal{L}_\mathcal{F}$ are used to form a total loss function with associated weights which are minimized.

3.1 Physics Informed Neural Networks

In this study we focus on finding ODE solutions using a machine learning approaches. We introduce a neural network to approximate a desired solution $x(t)$ of the ODE, with $\hat{x}_\theta(t) \approx x(t)$ where $\theta = (W, b)$ is a set of model parameters, (weight matrices and bias vectors) of the network.

The simplest of these forms is the forward neural network, which is also known as a is the feedforward neural network. Consider an L -hidden-layer fully-connected neural network with hidden layers of width N_l for layer l and pointwise nonlinearities ϕ . Let $x \in \mathbb{R}^{d_{in}}$ denote the input to the network, and let $z^{(l)} \in \mathbb{R}^{d_{out}}$ denote its output. The i -th component of the activations in the l -th layer, post-nonlinearity and post affine transformation, are denoted $h_i^{(l)}$ and $z_i^{(l)}$ respectively, and let $h_i^{(0)}(x) = x_i$ for the input initial and

$$z_i^{(0)} = b_i^{(0)} + \sum_{k=1}^{d_{in}} W_{ik}^{(0)} h_k^{(0)}(x) \tag{9}$$

Weight and bias parameters for the l -th layer are denote by $W_{ij}^{(l)}$ and $b_i^{(l)}$ respectively. The i -th component of the network output, $z_i^{(l)}$, is computed as,

$$z_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{N_l} W_{ij}^{(1)} h_j^{(1)}(x), \quad h_j^{(1)}(x) = \phi \left[b_j^{(0)} + \sum_{k=1}^{d_{in}} W_{kj}^{(0)} x_k \right] \tag{10}$$

Denoting the output of hidden layers by $h_i^{(l)}$, where $l \in \{1, \dots, L\}$, $i \in \{1, \dots, N_l\}$ denote the indices of the neuron within the layer that receives the information from the neurons of the previous layer $h_j^{(l-1)}$, $j \in \{1, \dots, N_{l-1}\}$. The output j represented by the j -th neuron in the output layer, is connected to the input vector x via a biased weighted sum and an non-linear activation function ϕ , which is applied element-wisely, the most commonly used hyperbolic tangent \tanh function. The j -th component of the network output, $h_j^{(l)}$, is computed as:

$$h_j^{(l)}(x) = \phi \left(z_i^{(l-1)}(x) \right) \tag{11}$$

where

$$z_i^{(l)}(x) = b_i^{(l)} + \sum_{j=1}^{N_{l-1}} W_{ij}^{(l)} h_j^{(l-1)}(x) \tag{12}$$

For convenience, the parameters of the neural network are combined into a vector of parameters $\theta = (b^{(0)}, W^{(0)}, b^{(1)}, W^{(1)}, \dots, b^{(L)}, W^{(L)})$ and input data $x^{(0)} = (1, x_1, \dots, x_{N_l})$. The computation for a network with L hidden layers is:

$$\hat{x}_\theta(t) = \hat{f} \left[h^{(L)} \left(z^{(L)} \left(h^{(L-1)} \left(\dots z^{(2)} \left(h^{(2)} \left(z^{(1)} \left(h^{(1)} \left(z^{(0)} \right) \right) \right) \right) \right) \right) \right) \right] \tag{13}$$

where \hat{f} can be seen as the identity function. When L is large it is called a deep neural network, and each pre-activation function $z^{(L)}(x)$ is typically a linear

operation with matrix $W^{(L)}$ and bias $b^{(L)}$, which can be combined with the parameters θ .

The goal is to calibrate its parameters $\theta = \{W^{(l)}, b^{(l)}\}_{l=1}^L$ such that $\hat{x}_\theta(t)$ approximates the target solution $x_\theta(t)$.

The resolution of the ODE is reduced to an optimization problem as follows, [20], suppose we have a set of N_d data is available for the known solution at different times $t_d^{(i)}$, i.e.

$$\left\{ t_d^{(i)}, x_d^{(i)}, \hat{x}_d^{(i)} \right\}_{i=1}^{N_d} \tag{14}$$

that are the training data, which include the initial condition. The optimization problem is based on the minimization of a loss function that can be expressed as,

$$\mathcal{L}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} \|\hat{x}_\theta(t_i) - x_d^{(i)}\|^2 \tag{15}$$

In a similar way, defining a loss function \mathcal{L}_{iv} corresponding to the knowledge of the initial condition,

$$\mathcal{L}_b = \frac{1}{N_b} \sum_{i=1}^{N_b} \|\hat{x}_\theta(t_i) - x_b^{(i)}\|^2 \tag{16}$$

where a set of N_b known data is imposed via

$$\left\{ t_b^{(i)}, x_b^{(i)}, \hat{x}_b^{(i)} \right\}_{i=1}^{N_b} \tag{17}$$

Finally, another loss function can be also obtained as,

$$\mathcal{L}_\mathcal{F} = \frac{1}{N_c} \sum_{i=1}^{N_d} \|\mathcal{F}(\hat{x}_\theta(t_i))\|^2 \tag{18}$$

that must be evaluated on a set of N_c data points, generally called collocation points, see [20] for details, where N_c which are not necessarily coinciding in time values with N_d . The literature recommends using automatic differentiation to compute derivatives with respect to the network weights. A composite loss function is generally formed as

$$\mathcal{L}(\theta) = \omega_d \mathcal{L}_d(\theta) + \omega_b \mathcal{L}_b(\theta) + \omega_\mathcal{F} \mathcal{L}_\mathcal{F} \tag{19}$$

The appropriate choice of hyperparameters $(\omega_d, \omega_b, \omega_\mathcal{F})$ allows optimizing the general loss function based on the partial losses. A gradient descent algorithm is used until convergence towards the minimum is obtained

$$\theta^{(i+1)} = \theta^{(i)} - \eta \nabla_\theta \mathcal{L}(\theta^{(i)}) \tag{20}$$

3.2 Probabilistic Solutions to ODE

Consider the initial value problem defined by the equation (1), and we are interested in estimating the solutions states of the continuous system using a Gaussian process. A Gaussian Process (GP) is a collection of random variables with the property that the joint distribution of any of its subset is joint Gaussian distribution. GP have some advantages because to known well-founded mathematical properties, which ensures confidence in modeling the prediction, there is detailed explanation in [6, 25, 26].

In real-world systems, the solution states in (1) are only partially observed. Now the continuous dynamical system can be written in the form:

$$\begin{aligned} \frac{dx}{dt} &= f(x(t), \theta), \quad x(t_0) = x_0 \\ y(t) &= h(x(t)) + \epsilon(t), \quad \epsilon(t) \sim N(0, \sigma_\epsilon^2) \end{aligned} \quad (21)$$

We wish to recover the latent function $x(t)$. In terms of a discrete-time stochastic dynamical system

$$\begin{aligned} x_t &= f(x_{t-1}) + \nu_{t-1}, \quad x_0 \sim p(\cdot), \quad \nu_{k-1} \sim N(0, Q) \\ y_t &= h(x_t) + \epsilon_{t-1}, \quad \epsilon_t \sim N(0, R) \end{aligned} \quad (22)$$

We formulate the approximation at discrete points as a Bayesian inference problem, where: $y_{1:t} = (y_1, y_2, \dots, y_t)$ are the observed data, $x_{1:t} = (x_1, x_2, \dots, x_t)$ the unknown states, $L(y_{1:t}|x_{1:t})$ is the likelihood, $P(x_{1:t})$ is the prior distribution over the solution space, and $P(x_{1:t}|y_{1:t})$ posterior distribution is obtained by:

$$P(x_{1:t}|y_{1:t}) = \frac{L(y_{1:t}|x_{1:t})P(x_{1:t})}{\int L(y_{1:t}|x_{1:t})P(x_{1:t})dx_{1:t}} \quad (23)$$

In this section, we use a Gaussian Process (GP) to find probabilistically the states solutions of the ODE. Assume we have $y_{1:n} = (y_1, y_2, \dots, y_n)^T$ as the values of the target function, where $y_i = f(x_i) \in \mathbb{R}$ are observations at locations $x_{1:n} = (x_1, \dots, x_n)^T$ where x_i are d-dimensional vectors in the domain $\mathcal{D} \subseteq \mathbb{R}^d$. We aim to use a GP $f(\cdot, \cdot) : \mathcal{D} \times \Omega \rightarrow \mathbb{R}$, to approximate the underlying target function. Typically, $y = f(x)$ is denoted as

$$f(x) \sim GP(m(x), k(x, x')) \quad (24)$$

where $m(\cdot) : \mathcal{D} \rightarrow \mathbb{R}$ is parametric fuction and $k(\cdot, \cdot) : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ admits a positive semi-definite matrix, defined by:

$$m(x) = \mathbb{E}[f(x)], \quad k(x_i, x_j) = \mathbb{E}\{[f(x_i) - m(x_i)][f(x_j) - m(x_j)]\} \quad (25)$$

GPs generalize the Gaussian distribution, are of infinite dimension and can be considered as a non-parametric method, and are defined in terms of the mean and covariance function, [6].

Consider the GP regression problem with input-output training pairs $\{(x_i, y_i)\}_{i=1}^n$:

$$f(x) \sim GP(\mathbf{0}, k(x, x')) \tag{26}$$

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{27}$$

A widely used kernel is the standard squared exponential covariance kernel with an additive independent identically distributed Gaussian noise term ϵ with variance σ^2 :

$$k(x, x') = \alpha^2 \exp\left(-\frac{\|x - x'\|_2^2}{2l}\right) + \sigma^2 \delta_{x,x'} \tag{28}$$

where $\delta_{x,x'}$ is a Kronecker delta function, l is the length-scale, and α^2 is the signal variance. In general, by assuming zero mean function $m(x) = 0$, we use $\theta = (\alpha, l, \sigma)$ to denote the hyperparameters, and they are determined based on the training data.

The joint distribution over test and training points is:

$$\begin{pmatrix} f \\ f_{new} \end{pmatrix} \sim GP\left(\begin{pmatrix} m(x) \\ m(x_{new}) \end{pmatrix}; \begin{pmatrix} K(x, x) + \sigma_\epsilon^2 \mathbf{I} & K(x, x_{new}) \\ K(x_{new}, x) & K(x_{new}, x_{new}) + \sigma_\epsilon^2 \mathbf{I} \end{pmatrix}\right) \tag{29}$$

Thus, we can rewrite f_{new} as a conditional:

$$f_{new} | f, x, y \sim N\left(m(x_{new})^{post}, K^{post}(f(x_{new}), f(x_{new}))\right) \tag{30}$$

The posterior mean for an unseen test input x_{new} is given by

$$m^{post}(x_{new}) = m(x_{new}) + K(x_{new}, x) (K(x, x) + \sigma_\epsilon^2 \mathbf{I})^{-1} [f - m(x)] \tag{31}$$

The posterior variance for an unseen test input x_{new} is given by

$$K^{post} = K(x_{new}, x_{new}) - K(x_{new}, \mathbf{x}) (K(\mathbf{x}, \mathbf{x}) + \sigma_\epsilon^2 \mathbf{I})^{-1} K(\mathbf{x}, x_{new}) + \sigma_\epsilon^2 \mathbf{I} \tag{32}$$

where

$$K(\mathbf{x}, \mathbf{x}) = \begin{pmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{pmatrix}$$

and

$$\mathbf{x} = (x_1, \dots, x_n)$$

4 Results

In this section, the methodology is shown using two synthetic models. In our simulation studies, we demonstrate the performance of the previous approaches on two dynamical systems: the Lotka Volterra model and the FitzHugh Nagumo

model by comparing the estimated parameters and state variables with their true values. Additionally, real COVID-19 data was used to infer the parameters of the SIR model.

The parameters for generating the simulations are as follows: for training the PINNs, we used 4 hidden layers with 32 neurons each for 12,000 epochs, and 5 hidden layers with 32 neurons each for 24,000 epochs, for $\sigma = 0.1$ and $\sigma = 0.5$, respectively. The learning rate in both cases was 3×10^{-3} . For the method combining probabilistic solutions and the Hamiltonian Monte Carlo algorithm, we set `draws` = 1000, `tune` = 1000, `target_accept` = 0.9, `chains` = 4, and `cores` = 4. Here, `draws` specifies the number of samples, `tune` indicates the number of tuning steps, `target_accept` defines the acceptance probability, and `chains` and `cores` represent the number of independent chains and CPU cores used, respectively. To evaluate the efficiency of the methods in state-parameter inference, we used the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

4.1 Simulation Lotka-Volterra Model

The Lotka-Volterra model, also known as the predator-prey equations, is a mathematical framework for understanding the interactions between predators and their prey in an ecosystem. It uses nonlinear differential equations to describe the cyclic dynamics where predator populations rise and fall in response to prey availability. This model, a specific instance of the broader Kolmogorov model, helps predict predator-prey interactions across various environments, making it essential in ecological modeling [29].

$$\frac{dS(t)}{dt} = \alpha S(t) - \beta S(t)W(t) + \sigma \quad (33)$$

$$\frac{dW(t)}{dt} = \delta S(t)W(t) - \gamma W(t) + \sigma \quad (34)$$

where $\theta = [\alpha, \beta, \gamma, \delta]^T$ and $\mathbf{X} = [S, W]$. The observed data are generated using Runge-Kutta method over the interval $[0, 5]$ with $\theta = [2, 1, 4, 1]$ an initial state values $S(0) = 5$ and $W(0) = 3$.

In Table 1, we compare the performance of different techniques for varying σ values. The results show the estimated parameters α , β , γ and δ , along with the computation time for each technique. Also, Table 2 presents a comparison of techniques for the functions $S(t)$ and $W(t)$ highlighting the RMSE and MAE values for each method.

Figure 1 illustrates the state-parameter inference of the Lotka-Volterra model (S, W) using probabilistic solutions together with the Hamiltonian Monte Carlo algorithm and Physics-Informed Neural Networks, respectively.

4.2 Simulation Fitzhugh-Nagumo Model

The Fitzhugh-Nagumo equations are a simplified version of the Hodgkin-Huxley model, which was developed to describe the initiation and propagation of electrical signals in the neuron of the giant squid axon. The Hodgkin-Huxley model,

Table 1. Comparison of Techniques for different σ values

σ	Technique	α	β	γ	δ	Time (s)
0.1	GP+HMC	2.014	1.006	3.981	0.997	36
	PINNs	1.991	1.001	4.083	1.031	155
0.5	GP+HMC	2.003	1.000	4.005	1.012	38
	PINNs	1.853	0.983	4.245	1.058	160

Table 2. Comparison of techniques for functions $S(t)$ and $W(t)$

Function	Technique	RMSE	MAE
$S(t)$	GP+HMC	0.044	0.041
	PINNs	0.106	0.081
$W(t)$	GP+HMC	0.019	0.016
	PINNs	0.14	0.122

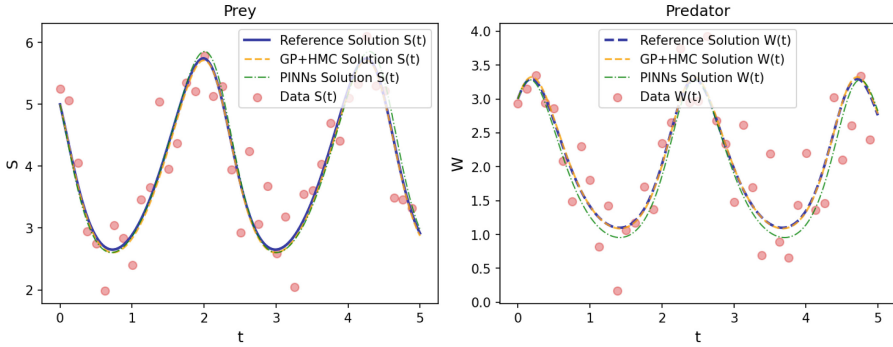


Fig. 1. The left graphic shows the inferred state of the prey ($S(t)$) using PINNs and GP + HMC compared to a reference solution. The right graphic displays the same comparison for the predator state ($W(t)$).

while detailed and accurate, is mathematically complex and computationally intensive due to its use of four coupled differential equations with nonlinear terms [28].

The Fitzhugh-Nagumo model simplifies this into just two coupled differential equations that capture the essential features of the neuron dynamics: the activation and the recovery variables. The activation variable represents the membrane potential, while the recovery variable accounts for the recovery process of the

membrane’s ion channels that are essential for the neuron to fire another action potential [27].

$$\frac{dV(t)}{dt} = \gamma \left(V(t) - \frac{V(t)^3}{3} \right) + R(t) + \sigma \tag{35}$$

$$\frac{dR(t)}{dt} = -(V(t) - \alpha + \beta R(t)) + \sigma \tag{36}$$

where $\theta = [\alpha, \beta, \gamma]^T$ and $\mathbf{X} = [V, R]$. The observed data are generated using Runge-Kutta method over the interval $[0, 5]$ with $\theta = [0.2, 0.1, 3]$ an initial state values $S(0) = 1$ and $W(0) = 3$.

In Table 3, we compare the performance of different techniques for varying σ values. The results show the estimated parameters α , β and γ along with the computation time for each technique. Also, Table 4 presents a comparison of techniques for the functions $R(t)$ and $V(t)$ highlighting the RMSE and MAE values for each method.

Table 3. Comparison of techniques for different σ values

σ	Technique	α	β	γ	Time
0.1	GP+HMC	0.149	0.101	2.994	38
	PINNs	0.204	0.121	3.042	378
0.5	GP+HMC	0.151	0.102	3.016	38
	PINNs	0.271	0.130	3.212	385

Table 4. Comparison of techniques for functions $V(t)$ and $R(t)$

Function	Technique	RMSE	MAE
$V(t)$	GP+HMC	0.13	0.059
	PINNs	0.633	0.357
$R(t)$	GP+HMC	0.106	0.089
	PINNs	0.516	0.369

Figure 2 illustrates the state-parameter inference of the Fitzhugh-Nagumo model (V, R) using probabilistic solutions together with the Hamiltonian Monte Carlo algorithm and Physics-Informed Neural Networks, respectively.

4.3 SIR Model for COVID Data

In this session, we show how our methodology works with real data. The COVID-19 dataset from Kaggle [30] includes key variables essential for modeling the

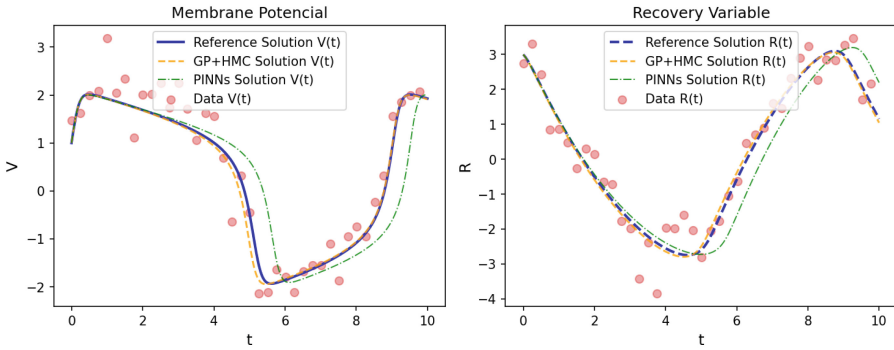


Fig. 2. The left graphic shows the inferred state of the Membrane Potential ($V(t)$) using PINNs and GP + HMC compared to a reference solution. The right graphic displays the same comparison for the Recovery Variable ($R(t)$) state.

spread of the virus using the SIR (Susceptible, Infected, Recovered) model, specifically for Japan in the year 2020 between August and December (Fig. 3). The “Susceptible” (S) variable represents the number of individuals who are at risk of contracting COVID-19, as they have not yet been infected and lack immunity. The “Infected” (I) variable indicates the current number of individuals who have contracted the virus and are capable of transmitting it to those in the susceptible category. Lastly, the “Recovered” (R) variable counts the individuals who have recovered from the infection and are presumed to have acquired immunity, thus no longer susceptible to reinfection. The results of our analysis are presented in Table 5 and Fig. 4. The SIR model is described as follows:

$$\frac{dS}{dt} = -\beta SI \tag{37}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{38}$$

$$\frac{dR}{dt} = \gamma I \tag{39}$$

where S , I , and R represent susceptible, infected, and recovered individuals, respectively. The parameters β and γ denote the transmission rate and recovery rate, respectively.

Table 5. Comparison of techniques for estimating parameters of SIR model with COVID-data.

Technique	β	γ	Time (s)
GP+HMC	0.0716	0.0632	55
PINNs	0.0705	0.0615	232

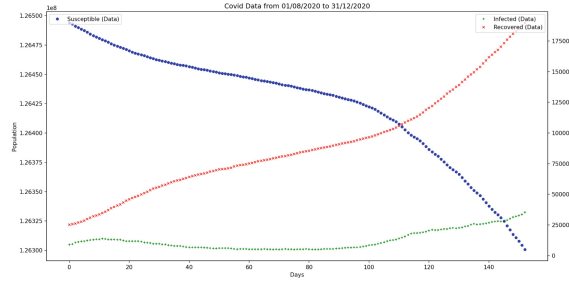


Fig. 3. COVID data from August, 01 2020 to December, 31 2020.

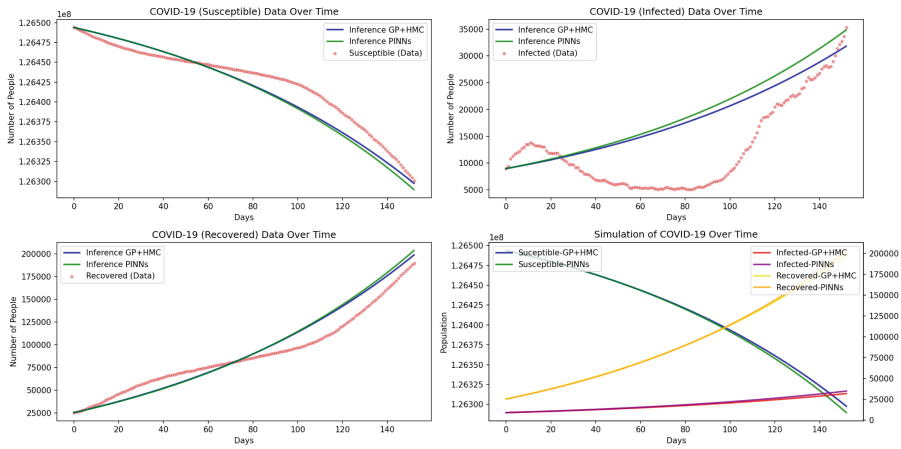


Fig. 4. The top left graphic shows the inferred state of the Susceptible (S) using PINNs and GP + HMC compared to a real data. The top right and bottom left of graphic displays the same comparison for the infected state (I) and the recovered state (R), respectively. Finally, on the bottom right of graphic shows the simulation of all states obtained by PINNs and GP+HMC.

5 Conclusions and Discussion

In this work, we present two novel non-parametric methods for estimating large-scale ODEs from noisy data with nonlinear structures. The first method encodes the physical laws of the studied dynamic systems through the design of DNN architectures, expressed in terms of function compositions. The main objective is to solve problems by optimizing under physical constraints, approximating the unknown function (solution states) using DNNs. Loss functions that penalize the training process are utilized, considering both the physical model and the feedback mechanism provided by the data. This method allows for the approximation of admissible solutions that respect the physical laws governing the sys-

tem's temporal dynamics without violating the principle of mass conservation. The second non-parametric method is probabilistic, belonging to the emerging paradigm of statistical learning. The main feature of probabilistic programming is based on Bayesian statistics, using conditional distributions on parameters given the observed data. This methodology becomes a powerful tool for constructing complex models in various scenarios. We use the PyMC library [9], a probabilistic programming platform that combines GP with samplers based on the HMC algorithm and automatic transformations of constrained random variables. This methodology facilitates the construction of flexible and accurate statistical models. Two simulation examples and a real data analysis have substantiated the validity and effectiveness of our methods. We have shown how our method allows us to estimate parameters and solution states in ODE systems, considering complex functional relationships. The metrics used show estimation errors with little variability. In conclusion, this methodology is a modern alternative way of estimating parameters and solutions of ODE systems, using flexible, robust, efficient and easy-to-interpret techniques.

Acknowledgements. This work was partly financed by the project Statistical Methods for Modeling data generated by Complex Systems REGINV- MATH23-06, Yachay Tech University.

References

1. Cuomo, S., Di Cola, V., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F.: Scientific machine learning through physics-informed neural networks: where we are and what's next. *J. Sci. Comput.* **93**(2), 1–88 (2022)
2. Baty, H., Baty, L.: Solving Differential Equations Using Physics Informed Deep Learning: A Hand-on Tutorial with Benchmark Tests, pp. 1–23. arXiv preprint [arXiv:2302.12260](https://arxiv.org/abs/2302.12260) (2023)
3. Winter, S., Campbell, T., Lin, L., Srivastava, S., Dunson, D.: Machine Learning and the Future of Bayesian Computation. arXiv preprint [arXiv:2304.11251](https://arxiv.org/abs/2304.11251) (2023)
4. Space-time error estimates for deep neural network approximations for differential equations. *Adv. Comput. Math.* **49**(1), 2–78 (2023)
5. Carnazza, F., Carollo, F., Andergassen, S., Martius, G., Klopotek, M., Lesanovsky, I.: Machine learning stochastic differential equations for the evolution of order parameters of classical many-body systems in and out of equilibrium. arXiv preprint [arXiv:2402.03913](https://arxiv.org/abs/2402.03913) (2024)
6. Christopher, W., Rasmussen, C.: *Gaussian Processes for Machine Learning*, vol. 2, no. 3. MIT Press, Cambridge (2006)
7. Barber, D., Wang, Y.: Gaussian processes for Bayesian estimation in ordinary differential equations. In: *International Conference on Machine Learning*, pp. 1485–1493 (2014)
8. Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., Sohl-Dickstein, J.: Deep neural networks as Gaussian processes. arXiv preprint [arXiv:1711.00165](https://arxiv.org/abs/1711.00165) (2017)
9. Salvatier, J., Wiecki, T., Fonnesbeck, C.: Probabilistic programming in python using pymc3. *Peer J Comput. Sci.* **2** (2016)





10. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222 (1987)
11. Neal, R.M.: MCMC using Hamiltonian dynamics. In: *Handbook of Markov Chain Monte Carlo* (Chap. 5) (2012)
12. Betancourt, M.: A conceptual introduction to Hamiltonian Monte Carlo. [arXiv:1701.02434](https://arxiv.org/abs/1701.02434) [stat.ME] (2017)
13. Owhadi, H.: Bayesian numerical homogenization. *Multiscale Model. Simul.* **13**(3), 812–828 (2015)
14. Raissia, M., Perdikaris, P., Karniadakis, G.E.: Inferring solutions of differential equations using noisy multi-fidelity data. *J. Comput. Phys.* **335**, 736–746 (2017)
15. Raissi, M.: Deep hidden physics models: deep learning of nonlinear partial differential equations. *J. Mach. Learn. Res.* **19**(25), 1–24 (2018)
16. Raissia, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019)
17. Baty, H.: A hands-on introduction to physics-informed neural networks for solving partial differential equations with benchmark tests taken from astrophysics and plasma physics. *arXiv preprint arXiv:2403.00599* (2024)
18. Le Cun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015). <https://doi.org/10.1038/nature14539>
19. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016). <http://www.deeplearningbook.org>
20. Baty, H.: Solving stiff ordinary differential equations using physics informed neural networks (PINNs): simple recipes to improve training of vanilla-PINNs, pp. 1–16. *arXiv preprint arXiv:2304.08289* (2023)
21. Kypriadis, A., Lagaris, I.E., Aristidis, L., Parsopoulos, K.: Optimized neural forms for solving ordinary differential equations. *arXiv preprint arXiv:2404.19454* (2024)
22. De Florio, M., Schiassi, E., Calabrò, F., Furfaro, R.: Physics-informed neural networks for 2nd order ODEs with sharp gradients. *J. Comput. Appl. Math.*
23. Garcia, P.: Modeling systems with machine learning based differential equations. *Chaos, Solitons & Fractals* **165**, 1–12 (2022)
24. Cedeño, N., Infante, S.: Estimation of ordinary differential equations solutions with Gaussian processes and polynomial chaos expansion. In: Salgado Guerrero, J.P., Chicaiza Espinosa, J., Cerrada Lozada, M., Berrezueta-Guzman, S. (eds.) *Information and Communication Technologies. TICEC 2021. Communications in Computer and Information Science*, vol. 1456 (2021)
25. Tronarp, F., Kersting, H., Särkkä, S., Hennig, P.: Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering. *Stat. Comput.* 1–19 (2019)
26. Román, K., Cumbicus, A., Infante, S., Fonseca-Delgado, R.: Deep Gaussian processes and infinite neural networks for the analysis of EEG signals in Alzheimer’s diseases. *Revista de Matemática: Teoría y Aplicaciones* **29**(2), 289–312 (2022)
27. Huang, H., Handel, A., Song, X.: A Bayesian approach to estimate parameters of ordinary differential equation. *Comput. Stat.* **35** (2020). <https://doi.org/10.1007/s00180-020-00962-8>
28. Benson, A.P., Stevenson-Cocks, H.J., Whittaker, D.G., White, E., Colman, M.A.: Multi-scale approaches for the simulation of cardiac electrophysiology: II—Tissue-level structure and function. *Methods* **185**, 60–81 (2021). *Methods on Simulation in Biomedicine*. ISSN 1046-202. <https://doi.org/10.1016/j.ymeth.2020.01.010>. <https://www.sciencedirect.com/science/article/pii/S1046202319302245>

29. Kek, S.L., Chen, C.Y., Chan, S.Q.: IOS Press ebooks—First-Order linear ordinary differential equation for regression modelling. <https://ebooks.iospress.nl/doi/10.3233/FAIA231184>
30. Guidotti, E.: A worldwide epidemiological database for COVID-19 at fine-grained spatial resolution. *Sci. Data* **9**(1), 112 (2022). <https://doi.org/10.1038/s41597-022-01245-1>

**IoT, Embedded Systems,
and Applications in Healthcare
and Industrial Environments**



Pre-processing of the Text of ECU 911 Emergency Calls

Marcos Orellana^(✉) , Pablo Andres Molina Pinos ,
Patricio Santiago García-Montero , and Jorge Luis Zambrano-Martinez 

Computer Science Research & Development Laboratory (LIDI), Universidad del
Azuay, Cuenca, Ecuador

{marore, jorge.zambrano}@uazuay.edu.ec,
{pmolina, santyg20}@es.uazuay.edu.ec

Abstract. The function of the Integrated Security Service (ECU 911) is to receive emergency calls and coordinate the response through various emergency services. These interactions are recorded and stored as a backup of each call. The present research proposes the pre-processing of the text contained in said calls through the application of selected text formalization techniques, such as the use of dictionaries of idioms, consultation in the Dictionary of the Royal Spanish Academy (RAE), the use of spell checkers and Named Entity Recognition (NER). A bitmap is proposed as an auxiliary tool to facilitate decision-making and correct formalization during the formalization process. Finally, the score obtained by a semantic comparison model is evaluated, in which the informal transcribed text of the calls is compared with the formalized text to determine whether the formalization of text improves the performance of the Artificial Intelligence models. The results demonstrated that performance was improved or maintained in 74% of the test cases thanks to the combination of these techniques. .

Keywords: Text mining · Named Entity Recognition · Text Formalization · Natural Language Processing · Emergency calls

1 Introduction

In Ecuadorian territory, the Integrated Security Service (ECU 911) is the government entity responsible for providing an agile and effective response to dangerous emergencies. This institution's primary function is to receive auxiliary calls through the reserved telephone number 911 and coordinate the mobilization of the necessary resources, such as ambulances, fire departments, police units, and other emergency services, to respond to the needs promptly [14].

ECU 911 has a continuous monitoring system of surveillance cameras and other security devices deployed nationwide, which makes it possible to detect potential risk situations and alert the competent authorities [15].

Every call that reaches the ECU 911 call centre is meticulously recorded and stored, serving as a crucial backup for each emergency handled. However, these

records, while invaluable, pose a unique challenge. They are often expressed in colloquial language, with idioms, word repetitions, and other linguistic complexities. Those defects can be supported when Text-processing techniques transform informal language into a more structured and formal format. The effectiveness of this pre-processing is then evaluated to enhance and transform the performance of trained Artificial Intelligence (AI) models in Spanish.

The Spanish spoken in Ecuador is a treasure trove of linguistic diversity, with lexical and phonological variations that testify to the rich historical heritage and the unique influences of its regions: Coast, Sierra, Oriente, and Insular Region [4]. Due to this, handling colloquial expressions in Natural Language Processing (NLP) is not just a challenge but a complex and fascinating one. These linguistic phenomena demand specific strategies for their accurate formalization. Global dictionaries such as those of the Royal Spanish Academy (RAE) and specialized regional dictionaries or even manual dictionaries can be used if resources empower us to verify the existence and meaning of these expressions [9].

Informal or unstructured texts can be formalized with specific NLP techniques. These techniques are essential to understanding the transformation process and achieving the desired result. Applying various techniques offers the advantage of segmenting each stage of the formalization process, allowing precise control of the results in each phase.

Normalization algorithms are crucial in handling a wide range of Out-Of-Vocabulary (OOV) words, and correcting spelling errors. These algorithms involve identifying error patterns, such as expansions, abbreviations, shortenings, deletions, and omissions of punctuation marks, among others, to apply appropriate corrections. Spell checkers use specific algorithms; one of the best known is the Levenshtein distance, which measures the difference between two strings and their closest match to correct their spelling using a dictionary of known words [2].

In this context, the origin of the text is a crucial factor. The meaning of words and phrases can vary significantly depending on the geographical or cultural context. A term or expression can be understood differently for a region. For that, meticulous handling of idioms, regionalisms, and word repetitions is paramount.

Consequently, a coherent and well-structured final text using NLP techniques implies a more agile and precise response from the emergency service operators. It contributes with this valuable processed information in developing new technological tools for ECU 911

Finally, this document is organized as follows: Sect. 2 contains the related works, and Sect. 3 establishes the methodology of the work. Section 4 presents the results, and Sect. 5 is the discussion. Finally, Sect. 6 contains the conclusions of the work.

2 Related Work

Text pre-processing is an essential step for NLP. Pre-processing consists of converting raw or unstructured text into a format that a model can understand.

Primarily, it removes errors and anomalies from the text and transforms it into a formal format that is more straightforward for the model to understand. This literature review analyzes the main text pre-processing techniques proposed by several authors. The advantages and disadvantages of each procedure are discussed according to the characteristics of each data corpus used in each project.

Within the scope of text pre-processing, there are several alternatives. Thus, Moreno and Guirao [7] propose a method for pre-processing Spanish text corpora, emphasizing the classification of ambiguous and unknown words processed with *GRAMPAL*, a Spanish morphological processor. The modules were evaluated and presented a rate of 98.3% effectiveness when correctly labeling the words within their category.

On the other hand, Tessore et al. [17] propose different types of procedures for adequately transforming informal text into structured text. It ensures that the quality of transformation is directly proportional to the decrease in non-vocabulary words and a reduced number of tokens. They use Hunspell in the Spanish dictionary for spelling correction and morphological analysis. However, Ahn [1] proposes breaking down the task of extracting events into a series of classification sub-tasks, with each task being crucially handled by a machine-learned classifier. The classification sub-tasks are: i) Identification of anchors, ii) Identification of arguments, iii) Assignment of attributes, and iv) Co-reference of events.

Sheikha and Inkpen [16] propose the creation of a dictionary of common formal and informal words in a corpus, treating only the “informal” words to reduce the volume of processed data. The final coincidence percentage was 90%. Likewise, Wang and Kan [19] experimented with the recognition and segmentation of text extracted from blogs in China. They used the Peking University online dictionary as the formal lexicon and the list of informal words compiled from training instances as the informal lexicon. The results showed an efficiency percentage of correctly converting the text at 60%, noting that the rate is affected due to the large number of idioms and contextual interpretations typical of the language, which harms the system’s performance.

On the other hand, Zhixing et al. [18] propose a prototype using the K-Nearest Neighbors (KNN) algorithm to label words outside the vocabulary in a text extracted from Twitter conversations. Then, each word is transformed into a vector, and they use a classifier trained on ordinary words to label words outside the vocabulary. After labeling, the corpus is divided into regular expressions, single words, and multiple words. The possible fixes are grouped, and the option chosen to make the replacement is the one with the highest total score.

Alternatively, Fan et al. [6] propose a procedure for transforming unstructured data into a formal language. The procedure eliminates missing data or applies missing value imputation methods to replace these data with inferred values. The proposed methods include the KNN algorithm and regression models. The generated results demonstrate that the KNN algorithm can achieve satisfactory performance even with relatively high proportions of data.

Barreiro et al. [3] propose training a model called *esPERTo*, which learns from the root form of the language it is trained in. In this way, words can be identified and transformed more quickly into formal language. This model has the advantage of being able to train it with two or more variants of the language in question. In this way, the system analyzes each sentence, provides the formal alternative to each word that *esPERTo* has detected as informal, and replaces it. However, it is limited to identifying OOV words as accurately or as accurately as incomplete words as other models.

Despite the works mentioned above, there are no specific applications or techniques in the field of NLP for formalizing texts in Ecuadorian Spanish, especially the variation of the Spanish spoken in the southern region of Ecuador. This variety of Spanish, is one of the most characteristic dialects in the country. Furthermore, it is distinguished by its vocabulary, which includes colloquial expressions, idioms, regionalisms, and particular intonation [10].

3 Materials and Methods

In this study, we propose to formalize the conversations of the ECU 911, facilitated by Computer Science Research and Development Laboratory (LIDI) at Universidad del Azuay. The Systems Process Engineering Metamodel 2.0 (SPEM 2.0) specification was adopted to structure the research phases. In this way, the methodology was divided into four main tasks: i) Understanding of the data, ii) Pre-processing of data, iii) Formalization of texts, and iv) Validation of the Formalization. Additionally, Python was used throughout the development of this study. Figure 1 illustrates the diagram that represents the methodology to follow.

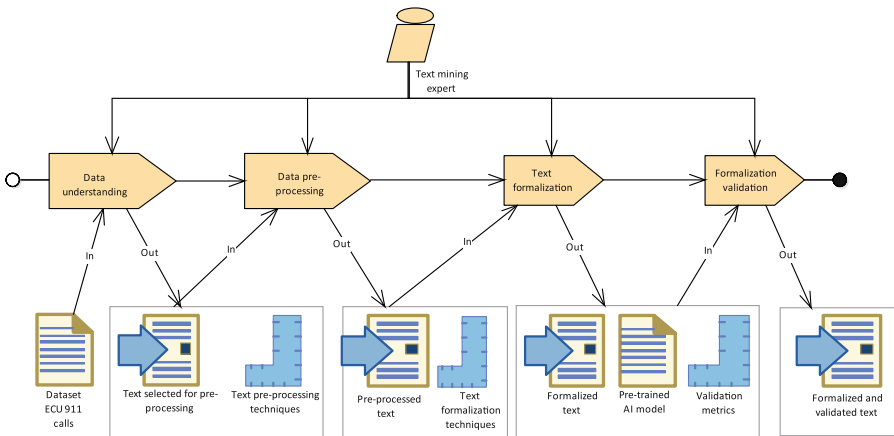


Fig. 1. Research Methodology Phases represented with SPEM 2.0

Table 1. Data set structure

Characteristic	Content
TRA_ID	Unique numeric identifier for each call.
ATA_TEXTO	Text of each line of a call.
ATA_SECUENCIA	Sequential numbering of lines within each call, reset to 1 when starting a new call.
ATA_ACTOR	Indicates whether the line belongs to an interaction between the Operator (“O”) or the Alerter (“A”).
INCIDENT_GRADE_NAME	The emergency level of the call is classified into four categories: “GREEN KEY”, “YELLOW KEY”, “ORANGE KEY”, and “RED KEY”, according to the parameters established in the ECU 911.

3.1 Data Understanding

The ECU 911 covers the entire national territory through several centers, operational rooms, and zonal coordination. In this study, the transcribed calls belong to the Zonal Coordination six (6), which includes the provinces of Cañar, Azuay, and Morona Santiago; both the provinces of Cañar and Azuay belong to the Sierra region, while the province of Morona Santiago belongs to the Eastern region [13]. In this phase, the data set of 1,000 calls transcribed for the formalization process was analyzed; its structure is presented in Table 1.

3.2 Data Pre-processing

The following techniques were applied for data pre-processing: Conversion to Lowercase, Tokenization, and Indexing. This series of techniques standardizes the information and ensures the reconstruction of the original format of the conversations after formalization.[8]

Conversion to Lowercase In this phase, all characters in the *ATA_TEXT* column were initially converted to lowercase using the *lower()* function. This standardization facilitates subsequent formalization processes by unifying the format of the words.

Tokenization This technique is a fundamental NLP process that decomposes text into smaller units called tokens. These tokens can be whole words, subwords, phrases, or even punctuation marks, depending on the type of Tokenization and the specific task. The rules that define how the text is divided are usually established in a previous stage called pre-tokenization, where the creation of tokens is restricted or enforced (e.g., the division of the text or corpus into white spaces). Additionally, tokens can be numerical representations used as input in statistical models or dense vectors called embeddings, used in modern architectures,

such as Bidirectional Encoder Representations from Transformers (BERT) or Generative Pre-trained Transformers (GPT)[12].

A Tokenization process was performed by dividing the text into smaller units (tokens). The *split()* function was employed to process the call transcriptions efficiently. This function leverages the inherent whitespace separation between words, acting as a delimiter to isolate individual words (tokens) within the transcript.

Indexación Each extracted token was then placed on a separate row in a new column named *WORDS_LIST*. Additionally, new columns were created for each unique word; an *ID* for identification purposes and another column to specify the *actor* associated with each word. This was achieved by concatenating the word with the corresponding values of the *ATA_SEQUENCE* and *ATA_ACTOR* columns to obtain the final identifier of each word, that is, its order within the new column. With this indexing process, it was possible to relate each word to the actor who pronounced it and their position in the call. This information is crucial to reconstruct the original format of the conversations after formalization.

3.3 Texts Formalization

Four techniques to formalize the text were applied in this phase: idiom dictionary, RAE search, Named-Entity Recognition (NER), and a spell checker supporting the Spanish language. These techniques were implemented in an order defined by a bitmap.

Idioms Dictionary. The idioms dictionary was prepared manually based on the “Diccionario de la real lengua morlaca” [5]. With this dictionary, an iterative process was implemented to loop through all the words in the text and search for matches in the idiom dictionary. The words found were replaced by their formal equivalents.

Search in the RAE. Python library *pyrae* was used to verify the correct spelling and existence of the words in the Spanish language. *dle()* method in that library allows you to search for words in the RAE dictionary and return their meaning if found or an error message if they do not exist. This method consulted each word on the list to determine its presence in the RAE.

NER. This is responsible for identifying and classifying named entities present in a text, such as names of people, organizations, and locations. This technique is beneficial for retrieving relevant information from a text and obtaining a clear view of names, locations, countries, or organizations present in the text [11].

This technique was also implemented because they are unique contextual elements and require special treatment. For this, the *spacy* Python library was used, with the *es_core_news_lg* model trained to identify entities named in Spanish. Table 2 shows the classification labels offered by the model.

Table 2. Label classification within the POS tagging model

Label	Description
['PER']	Person.
['LOC']	Location.
['ORG']	Organization.
['MISC']	Various entities.

Spell Checker. A spell checker was used using the *spellChecker* Python library. The *Spellchecker(language='es')* method provides a specific checker for the Spanish language. This method analyzes each word and selects, from several candidates, the most appropriate spelling correction using the Levensthein distance, which calculates the distance between words and determines the closest permutations to the original word.

After defining the guidelines for each technique, the challenge arose of determining the optimal order of application in each method. A bitmap was generated by applying the five techniques to the *WORDS_LIST* column to establish an application sequence and optimize the formalization results.

Bitmap. In the case of the idiom dictionary, “0” was assigned if the word was not an idiom, and a “1” if it was an idiom and has been replaced by its formal equivalent. For the RAE method, a “0” was assigned if the word was not found in the dictionary and a “1” if it was present. In NER, a “0” was assigned if the word was not a named entity and a “1” if it was. For the spell checker, a “0” was assigned if the word did not require correction and a “1” if it needed to be corrected.

3.4 Formalization Validation

At this stage, a semantic comparison was implemented to validate the formalization of the text. A sample of 20% was selected for a formal and informal semantic comparison. In this way, both comparisons’ semantic similarity changes are evaluated, and the cases in which the similarity index demonstrates improvements are determined. For an equitable comparison, the sample was filtered by considering the number of words in each and using the first standard deviation as a criterion.

The pretrained Sentence-Transformer model in Spanish *sentence_similarity_spanish_es* available in the Hugging Face Hub repository was used for the semantic comparison. With this model, the semantic comparison was performed using the cosine similarity metric between the vectors (embeddings) generated by the model. Two comparisons were made: informal text vs. informal text and informal text vs. formalized text, generating a column of scores for each possible combination within the 20% sample and selecting only the highest score from each call.

Table 3. The prepared idioms dictionary samples

Idioms	Formal word
“abombar”	“fastidiar” (“disturb”)
“avispado”	“astuto” (“cunning”)
“ayau”	“dolor” (“pain”).
“guagua”	“niño” (“kid”).
“ñuto”	“pequeño” (“little”)

Finally, a column with the final score was created by subtracting the scores of the two comparisons. This column generated three categories: positive, neutral, and negative. These categories allow the interpretation of the final formalization results. A positive score implies that the formalization process improved performance.

4 Results

For greater understanding, the results of this research were divided into three subsections. First, the results when applying the pre-processing techniques to the set of transcribed emergency call records; second, the results of the text formalization stage; and finally, the results of the validation of the formalization applied to the records.

4.1 Data Pre-processing

During the data pre-processing stage, the *INCIDENT_GRADE_NAME* column was discarded because it was irrelevant to the study. After the Tokenization process, 19,708 records in the *ATA_TEXT* column were transformed to 173,905 records in the new *WORDS_LIST* column.

4.2 Text Formalization

Idioms Dictionary. After the techniques were executed in parallel, the bitmap was generated. The results of each row were concatenated to obtain the resulting methods. Furthermore, these idioms were recorded with their formal language equivalent. Table 3 shows an example of the structure of this dictionary.

Bitmap. After the parallel execution of the techniques, the bitmap was generated. The results of each row were concatenated to obtain the resulting combinations of the methods. Table 4 shows the frequency distribution of these combinations in the bitmap.

Table 4 was used as a decision-making tool regarding the order in which the techniques were applied. To ensure linguistic accuracy, the idiom substitution technique was implemented first. By having a specialized dictionary as

Table 4. Frequency of combinations resulting from the bitmap

Combinations	Frequency
[1111]	27
[1110]	46
[1101]	23
[1100]	44
[1011]	12
[1001]	11
[0111]	1691
[0110]	12312
[0101]	758
[0100]	141123
[0011]	4282
[0010]	4818
[0001]	1463
[0000]	7295

a reference, the terms replaced by their formal equivalent safeguard the text's coherence. Therefore, using the bitmap in Table 4, a visual representation of the data where each bit represents a specific linguistic feature, all the records that contain the value of "1" in their first bit were selected; the idiom replacement technique was used in these records. Subsequently, the records that contain the value of "1" in their third bit and the value of "0" in their first and second bits were selected to apply the NER identification technique. Since, in many cases, they correspond to names, places, sectors, etc., they should not be replaced or corrected. Next, the records that contained the value of "1" in their fourth bit and the value of "0" in the rest of the bits were selected for the spelling correction technique. Finally, the records with the value of "0" in all their bits did not fit into any formalization technique.

Additionally, Fig. 2 shows the distribution of applied techniques to each text token.

These results demonstrated that the search for words in the RAE was the most used technique throughout the formalization process, covering the most significant number of words or tokens. This result indicates that the RAE recognizes the majority of words present in the calls and, therefore, are grammatically correct.

4.3 Formalization Validation

After data division, 200 records of the initial 1,000 were filtered, retaining only those with several words between 65 and 135, according to the criterion established by the first standard deviation. The total number of possible combinations of the 200 calls was 19,900 records. However, only the most significant were

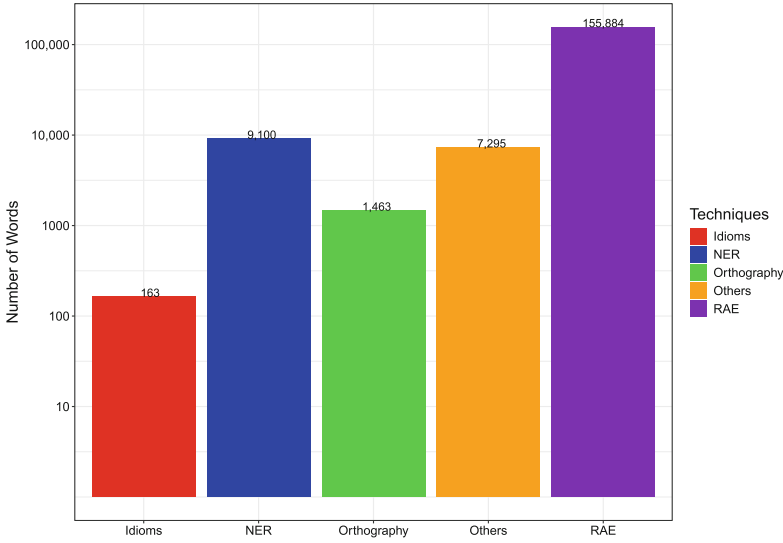


Fig. 2. Formalization techniques application in the words of the text.

selected to maintain the same number of records (200) and facilitate the interpretation of results. After classifying the scores into the established categories, it was determined that, of the 200 calls analyzed, the formalization improved the semantic comparison score in 73 cases, did not generate changes in 75 cases, and decreased the performance of the model in 52 cases, this distribution of presented in Fig. 3.

On the other hand, using the Gephi tool, all model performance score values were analyzed. The positive category increased from 36.5% to 40.11%, the neutral category decreased from 37.5% to 30.71%, and the negative category increased from 26% to 29.18%. However, the sum of the positive and neutral categories decreased slightly to 70.82%. This finding indicates that of the 19,900 possible model combinations, formalization positively impacted or maintained model performance in 14,093 (approximately 7 out of 10).

Positive and Neutral Score Representation. Within the group of calls where formalization improved the results, the degree of improvement in semantic comparison is an essential factor to consider. Figure 4 illustrates the number of calls and the percentage of improvement obtained after applying the formalization process proposed in this methodology.

Negative Score Representation. Likewise, analyzing the degree of reduction in semantic comparison is relevant in the group of calls where formalization decreased the results. Figure 5 illustrates the number of calls and the percentage reduction observed after applying the formalization process proposed in this methodology.

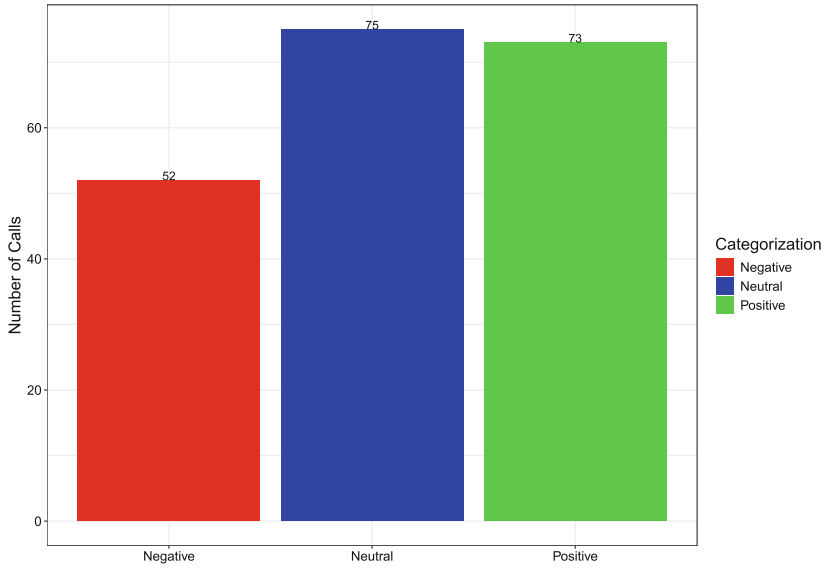


Fig. 3. Formalization process result.

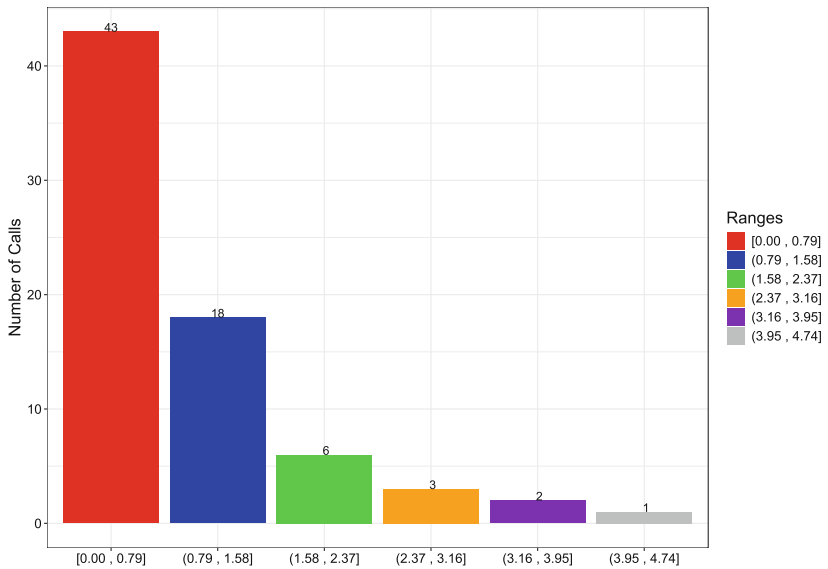


Fig. 4. Calls with positive score.

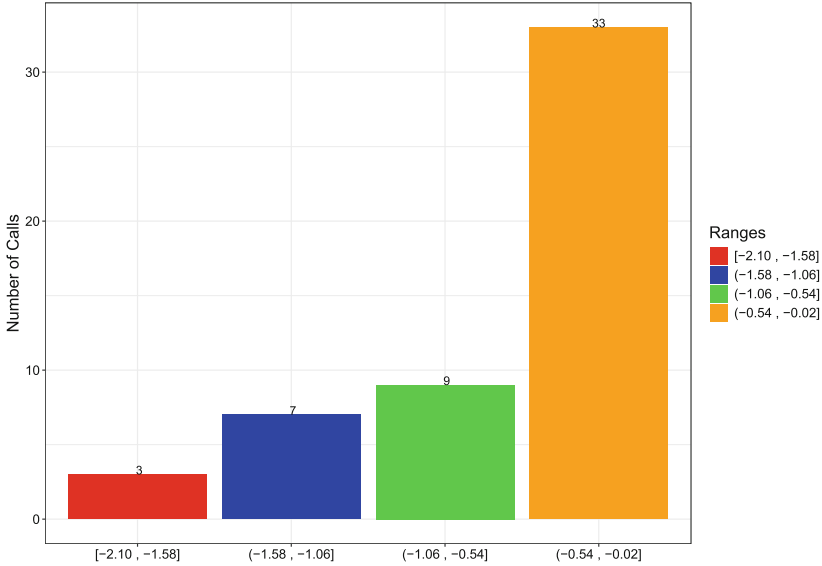


Fig. 5. Calls with negative score.

5 Discussion

5.1 Applied Techniques Effectiveness

Although the search in the RAE was the most used formalization technique, the contribution of the other techniques, although used to a lesser extent, was equally relevant. Consequently, it underlines the importance of employing a diverse approach that combines multiple techniques to address text formalization in this context. The presence of the “Others” category shown in the Fig. 2, composed chiefly of numerical data (such as identity cards and vehicle license plates), demonstrates the particularity of the language used in emergency calls, characterized by communication often imperfect and the inclusion of data that cannot be formalized.

5.2 Impact of Formalization on Model Performance

Formalization had a positive or neutral effect on most calls (74%), indicating an overall improvement in text quality. However, a significant percentage of calls (26%) showed decreased semantic performance after formalization. Detailed analysis of the improvement scores reveals that most improvements are concentrated in moderate ranges (0.00 to 0.79), as we can observe in Fig. 4. This suggests that the semantic comparison model, due to its pre-training, already essentially understands the context of the language used in the calls, even in its informal form.

An alternative to increase positive cases is to expand the idioms dictionary used in the replacement technique. A more complete dictionary, including expressions from other regions and cities, can improve semantic performance.

6 Conclusion

In this study, text pre-processing techniques were delved into, which made it possible to select and apply effective formalization methods to ECU 911 emergency calls. Thanks to tools such as bitmaps, idiom dictionaries, search in the RAE, spelling correction, and NER method, the text was formalized in a logical order.

The results have demonstrated the effectiveness of the formalization process. This demonstrates improved semantic performance in most calls. However, the percentage of negative cases due to language complexity indicates the need to refine formalization techniques further to optimize the performance of the semantic comparison model. Validation using this model confirmed an improvement in the formalization process. It supported the proposed methodology, demonstrating that text pre-processing significantly improves the models' performance compared to informal text.

Finally, it is proposed to evaluate the formalization with other semantic comparison models, which could provide a more comprehensive understanding of its effectiveness. Additionally, expanding dictionaries of idioms could enhance the accuracy of the formalization process. Testing different spelling checkers and applying the methodology to other data sources provides valuable insights for performing exhaustive comparisons and refining formalization techniques.

Acknowledgments. This work was supported by the vice rectorate of Research at Universidad del Azuay. Therefore, we thank them for their financial and academic support and the entire Computer Science Research & Development Laboratory (LIDI) staff.

References

1. Ahn, D.: The stages of event extraction. In: Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pp. 1–8 (2006)
2. Aliero, A.A., Bashir, S.A., Aliyu, H.O., Tafida, A.G., Bashar, U.K., Nasiru, M.D.: Systematic review on text normalization techniques and its approach to non-standard words. *Int. J. Comput. Appl.* (2023)
3. Barreiro, A., Rebelo-Arnold, I., Baptista, J., Mota, C., Garcez, I.: Automated paraphrasing of Portuguese informal into formal language. *LINGUAMATICA* **10**(2), 53–61 (2018)
4. Bolanos-Fabres, P.: *Expresiones populares y el contacto de lenguas* (2015)
5. Dirección Municipal de Relaciones Externas: *Diccionario de la real lengua morlaca* (2019)
6. Fan, C., Chen, M., Wang, X., Wang, J., Huang, B.: A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front. Energy Res.* **9**, 652801 (2021)

7. Moreno, A., Guirao, J.M.: Tagging a spontaneous speech corpus of Spanish. *Proceedings of RANLP-2003—Recent Advances in Natural Language Processing* (2003)
8. Orellana, M., García, P.S., Ramon, G.D., Zambrano-Martinez, J.L., Patiño-León, A., Serrano, M.V., Cedillo, P.: Semantic similarity of common verbal expressions in older adults through a pre-trained model. *Big Data Cogn. Comput.* **8**(1), 3 (2023)
9. Pastor, G.C.: Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In: *Fraseopragmática*, pp. 335–374. Frank & Timme (2013)
10. Portocarrero, A., Stewart, J.: A preliminary analysis of intonation patterns in Ecuadorian Cuencano Spanish. *Open Linguistics* **7**(1), 448–475 (2021)
11. Roy, A.: Recent trends in named entity recognition (NER). arXiv preprint [arXiv:2101.11420](https://arxiv.org/abs/2101.11420) (2021)
12. Schmidt, C.W., et al.: Tokenization is more than compression. arXiv preprint [arXiv:2402.18376](https://arxiv.org/abs/2402.18376) (2024)
13. Servicio Integrado de Seguridad ECU 911: Cobertura de los centros ecu 911. <https://www.ecu911.gob.ec/cobertura-nacional/> (2017)
14. Servicio Integrado de Seguridad ECU 911: Servicio integrado de seguridad ecu 911. <https://www.ecu911.gob.ec/> (2017)
15. Servicio Integrado de Seguridad ECU 911: ¿cómo solicitar un video del ecu 911? <https://www.ecu911.gob.ec/como-solicitar-un-video-del-ecu-911/> (2017)
16. Sheikha, F.A., Inkpen, D.: Generation of formal and informal sentences. In: *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 187–193 (2011)
17. Tessore, J.P., Esnaola, L.M., Russo, C.C., Baldassarri, S.: Comparative analysis of preprocessing tasks over social media texts in Spanish. In: *Proceedings of the XX International Conference on Human Computer Interaction*, pp. 1–8 (2019)
18. THU, Z.L., UPC, X.C., THU, S.W., UPC, L.P., UPC, A.P., UPC, J.T., Date, C.D.: Early informal language structure extraction prototype. *The European Community's Seventh Framework Programme* (2012)
19. Wang, A., Kan, M.Y.: Mining informal language from Chinese microtext: joint word recognition and segmentation. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 731–741 (2013)



Secure BLE Communication Between Android Devices and Embedded Systems for IoT Applications

Raúl Armas  and Marlon Navia  

Technical University of Manabí, Portoviejo 130103, Ecuador
{rarmas4485, marlon.navia}@utm.edu.ec

Abstract. Data privacy is an important key in current Internet of Things (IoT) communications. In this work, a novel mechanism for secure Bluetooth Low Energy communications between Android and IoT devices is proposed, based on a literature review about vulnerabilities and potential attacks over this protocol. The vulnerabilities and attacks found in the literature were analyzed and, with this basis, a mechanism was proposed that includes a series of application-level security measures to ensure secure communication. This mechanism involves the generation, sharing, and renewal of both symmetric and asymmetric security keys, as well as time and authenticity control by applying timestamps, unique keys, among other measures. As a result, according to the methodology, the proposal provides resistance against 71.43% of the attacks found in the analyzed works. The remaining 28.57% of attacks, not covered by the present proposal, do not impact the privacy or authenticity of the transmitted data, ensuring secure communication.

Keywords: Bluetooth smart · secure Bluetooth · cypher · Authentication

1 Introduction

The Bluetooth Low Energy (BLE) wireless communication protocol, also known as Bluetooth Smart, is ideal for short-range IoT devices that need to transmit data wirelessly and have strict power consumption requirements. Common examples of IoT devices using BLE include environmental sensors, fitness trackers, environment-aware systems in facilities [1], home and personal security systems [2], and it even played a significant role in the Covid-19 pandemic, as several government applications utilized a BLE-based exposure notification system [3, 4]. This communication technology has become a popular choice for establishing links between IoT systems and mobile devices, so BLE is now included in almost all key platform devices, from smartphones to tablets and laptops, facilitating the integration and control of these systems through applications.

In the last five years, many Bluetooth devices have been dual-mode (Bluetooth Classic + BLE). However, due to the continuous growth in demand for connected electronic devices, the number of single-mode BLE devices is expected to triple in the next five

years. When considering both single-mode BLE devices and dual-mode devices, approximately 95% (around 6.6 billion) of all Bluetooth devices are projected to include BLE by the year 2026, with a compound annual growth rate of 25% [5].

Considering that sensitive data can be transmitted via BLE, such as vital signs monitoring devices [6], it is essential to ensure secure communication. However, ensuring security in BLE communication presents unique challenges due to the wireless nature of the technology and vulnerabilities associated with data transfer. Despite the well-known vulnerabilities and attacks to this protocol, the standard security mechanisms are usually ineffective [7].

The main objective of this work is to propose a security scheme for BLE communication between Android devices and embedded systems. For this aim, an updated literature review has been performed regarding recent vulnerabilities and potential attacks on the protocol, focusing on the Android mobile operating system, as this is the most popular and widely used worldwide, covering 71.81% of the market [8].

This work has two significant contributions. The first is an updated review and analysis of the vulnerabilities and attacks that BLE may suffer. The second is a proposal for a security mechanism to provide secure communications through the BLE protocol.

The rest of the article is structured as follows: Sect. 2 provides a brief review of related works focused on security in BLE. The following section details the methodology applied in this research. Then, Sect. 4 presents the results in two parts: the first part provides an analysis of vulnerabilities and attacks of the protocol based on the found literature about the subject; and then, the proposal presented in this research is introduced to provide a higher level of security for applications using BLE. The research concludes with the findings.

2 Related Works

The security concerning the BLE protocol has attracted the attention of many researchers over the years; some of the most recent have focused on the evolution of its security throughout its versions [9]. Others have dedicated themselves to exploring the consequences of their vulnerabilities [10]. Wang et al. [11] designed a technique to obtain the information transmitted in a BLE connection between a computer and a peripheral device: by imitating a low-capacity device, the process of negotiating encryption and authentication keys can be eliminated, imposing a new key with the peripheral BLE device and pass authentication without user involvement. As a result, it can extract BLE packets and read or write data stored on the devices.

On the other hand, Pallavi et al. [12] demonstrated possible attacks on BLE devices that can be executed using low-cost tools to carry out spoofing, man-in-the-middle attacks, and firmware attacks. Likewise, Kurt Peker et al. [13] tested several BLE consumer wearable devices, including heart rate devices, demonstrating the possibility of data leakage when these transmit, including the measurements made. Due to this, these authors suggest that BLE devices must include a label informing the security level implemented.

Some authors have been concerned about the evolution of BLE applications: Klieme et al. [14] propose using the BLE protocol as a means to implement a web authentication

system based on the FIDO2 standard, which uses a client device, such as a computer, and an authenticator device, such as a cell phone. This system allows user to log in to websites without using passwords but rather their biometric data. When the user wants to log in from the client, it communicates through BLE with the cell phone, which obtains the user's biometric data and sends it back. Subsequently the client communicates with the server using the FIDO2 standard.

Banani et al. [15] present a proposal for enhancing security in BLE communications. This proposal consists of a dynamic light-weight encryption algorithm, which uses a simple XOR function for data encryption in nodes. This algorithm requires three different keys for encryption process (master key, node secret key, and temporal key), and its authors claim that it causes low overload in BLE sensor nodes. This proposal is focused on systems composed of several nodes that send data to a server. Besides, it needs a master key to be distributed in all nodes where it is implemented.

Sun and Tian [16] proposed a mechanism that provides address privacy protection for BLE devices. Their proposal consists of an enhanced resolvable private address (RPA) mechanism which overcomes the weakness of the original RPA. Even though this mechanism has not been tested in either simulation or real devices, authors demonstrate mathematically its performance.

Furthermore, Yang et al. [17] published a comprehensive review of next-generation applications built with BLE, obstacles to its adoption in new application areas, and current solutions from academia and industry that further expand its capabilities.

3 Materials and Methods

The literature review's methodology was based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), although a detailed analysis was performed instead of a meta-analysis. The PICOC method was used for systematic literature reviews to define the scope of the current state-of-the-art literature review:

- **Population:** Bluetooth Low Energy, Bluetooth Smart.
- **Intervention:** Android.
- **Comparison:** Vulnerability, attack, encryption, security.
- **Outcome:** Not applicable.
- **Context:** Not applicable.

Based on the PICOC formulation, the following search string was generated: (“*Bluetooth Low Energy*” OR “*Bluetooth Smart*”) AND (“*Android*”) AND (“*attack*” OR “*encryption*” OR “*security*” OR “*vulnerability*”). The bibliographic databases used to obtain information resources were: Scopus, ACM Digital Library, IEEE Xplore, and Science Direct. Inclusion and exclusion criteria were defined to filter the contributions obtained from the databases (Table 1).

For measuring the quality of the selected articles and obtaining the best results for this research, a quality assessment protocol was developed based on the question: “*Does it refer to secure BLE communication on Android devices and other devices or discuss any communication vulnerability or attack on it?*”.

Only those publications that passed this quality question were considered for the bibliographic study.

Table 1. Inclusion and exclusion criteria.

Inclusion	Exclusion
Books	Editorials
Book chapters	Abstracts
Journal articles	Not available
Conference articles	Review articles
	Conference review
	Published before 2019

By applying the search string in the bibliographic databases defined in the methodology, 47 resources were obtained. Of these, 14 were duplicates, meaning they were found in two or more bibliographic databases. When evaluating these resources according to the inclusion and exclusion criteria, 14 were accepted. After applying the quality assessment, at the end of PRISMA methodology, only eight were chosen for analysis and as the basis for developing the final proposal (Fig. 1).

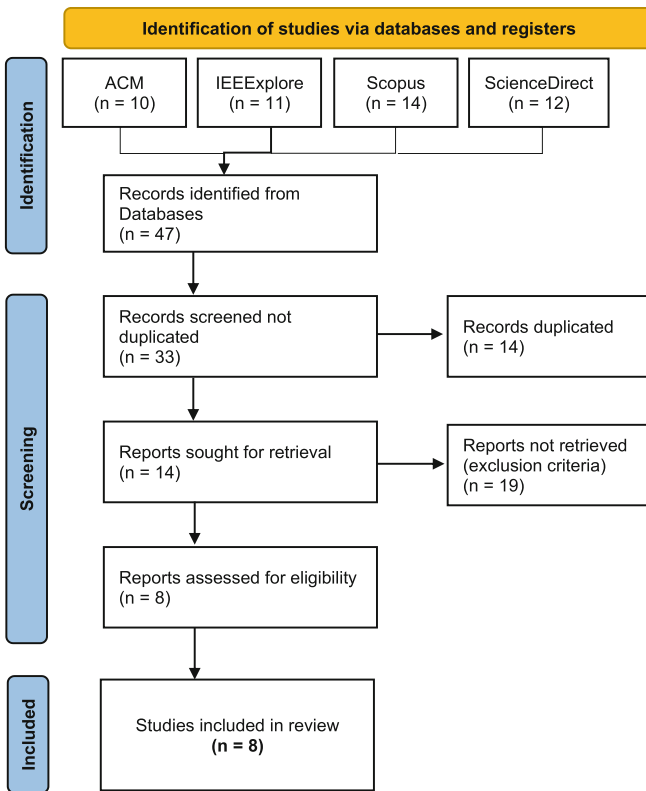


Fig. 1. PRISMA diagram search results.

A study of the attacks that BLE communication can suffer was conducted with the selected publications. An analysis was carried out for each of them, systematizing them with the vulnerabilities they exploit to execute through a cross matrix.

Based on this analysis, we proposed a mechanism to provide cyphered communication between Android and IoT devices to that avoid analyzed attacks.

4 Results

4.1 Found Vulnerabilities and Attacks

Table 2 shows references of selected publications and what was found in each one. It is worth mentioning that for the development of the final proposal, the official Bluetooth specification document [18] was also considered for reference.

Table 2. Publications selected for analysis.

Reference	What publication presents
[19]	Vulnerability that allows an application to bypass location restrictions
[20]	Data exfiltration by co-located applications
[21]	Impersonation attacks during the reconnection process
[22]	Link-level protocol vulnerabilities from version 4.2 onwards
[23]	Data exfiltration without the need for authentication or prior pairing
[24]	Relay attacks in government applications for COVID-19 exposure
[25]	Proposal for a security system based on BLE and HTTPS protocols
[26]	Vulnerabilities in Xiaomi's physical activity monitoring applications

BLE is designed to establish fast and backwards-compatible connections between devices, meaning two BLE devices of different versions can quickly establish communication. Due to its simple and low-power nature, advertising data such as attribute information and Media Access Control (MAC) address is always transmitted in plain text [18]. This is one of the principal vulnerabilities of BLE, as it allows an attacker attempting to connect to the device to obtain this information, clone its MAC address, and impersonate the device by advertising the same information.

Wu et al. [21] exploited this vulnerability to perform an **Impersonation Attack**. They found that Android devices run a reconnection procedure they termed Proactive Reconnection, in which they request to establish encryption after reconnecting. According to [18], if the encryption establishment process fails during the procedure, the client should either re-pair the connection with the server or abort it. However, the authors discovered that communication continues, making the attack possible (Fig. 2). This is due to a logical error in the implementation of BLE in the Android Application Programming Interface (API) which does not provide capabilities to detect encryption errors: if

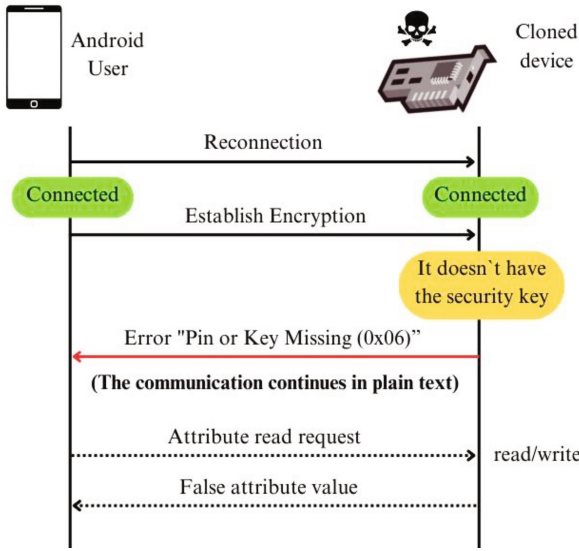


Fig. 2. Impersonation attack. If encryption establishment fails, communication can continue in plain text.

a user through their device connects with another and fails to establish encryption, the communication could continue in plain text without the user being aware.

Zhang et al. [22] delved deeper into the issue and discovered that communication continues in plain text if accessed attributes have read/write permissions, while if they have encrypted read/write permissions or encrypted and authenticated read/write permissions, Android automatically begins the re-pairing process (Table 3). If the attribute the attacker tries to access has encrypted or encrypted and authenticated permissions, the authors of [21] warn that the attacker can intercept this request and send a false value to avoid re-pairing.

Other attacks, besides the attacks mentioned earlier, involve the following:

Eavesdropping Attack: Zhang et al. [22] found that if communication can be downgraded to plain text using a fake device when the original devices reconnect, the communication will continue in plain text. This is because Android has already received error 0x06 and does not request encryption establishment. The authors of [21] referred to this type of reconnection, where the client device does not request encryption establishment, as Reactive Reconnection, and it is the default on Linux devices. Unlike the impersonation attack, in this case, the attacker does not communicate with the user but disconnects to eavesdrop on the user’s communication with the original device through a sniffer (Fig. 3).

Denial of Service (DDoS) Attack: The attacker forces re-pairing with the Android device from a fake device, modifying the permissions of the attributes to be encrypted or both encrypted and authenticated. When the user attempts to access any attributes, the fake device will send an error 0x0f or 0x05, compelling the Android device to re-pair

Table 3. Errors in BLE communication.

Error	Details	Reaction
Pin or Key Missing (0x06)	The peripheral device does not possess the security key	Android continues the communication in plain text
Insufficient Encryption (0x0f)	The security level of the link is not secure enough	Android automatically starts the re-pairing process
Insufficient Authentication (0x05)		

with the fake device automatically. This attack is possible by simulating a device with limited capabilities: starting from BLE version 4.2, Bluetooth SIG added four possible pairing methods based on the input and output capabilities of the devices:

- Just Works: No interaction is required.
- Passkey Entry: Requires entering a code on one device displayed on the screen of the other.
- Numeric Comparison: Requires confirming that both devices display the same code.
- OOB (Out Of Band): Requires a channel other than Bluetooth.

For the attacker to successfully pair the Android device with the fake device, they must force a Just Works pairing. To achieve this, the attacker intends, on the fake device, to have neither input nor output capabilities. This way, the pairing with the Android device occurs without the user's awareness. Finally, when the fake device is offline and the user connects to the original device, communication will come to a standstill because the Long-Term Key (LTK) changed in the re-pairing process. This state will continue until the user manually re-pairs the devices (Fig. 4). LTK is a shared cryptographic key used to establish a secure and encrypted connection between paired devices. This key is used as part of the authentication and encryption process to ensure the privacy and security of transmitted data. The process of generating and sharing the LTK typically occurs during the pairing process.

Man-in-the-Middle (MitM) Attack: To carry out this attack, the attacker captures the Identity Recognition Key (IRK) and the MAC address of the Android device through a re-pairing with a cloned device. Then, they use this information to generate a cloned device on the BLE network. Subsequently, the attacker connects the cloned devices with the originals. In this way, the attacker acts as a middleman in the communication, able to collect and modify data at will (Fig. 5). The IRK is used to generate resolvable private addresses in advertising and communication. A BLE device, such as a mobile phone, can be tracked if its MAC address is used in advertising and subsequent communication. BLE addresses this privacy issue through the IRK. Only a device with privacy requirements needs to distribute its IRK and actual MAC address to the peripheral device. If a mobile phone needs to protect its MAC address, it first distributes its IRK and MAC address to the peripheral device. Then, the mobile phone uses this IRK to generate a resolvable private address for its packets, and the peripheral device uses the mobile phone's IRK to resolve the private address.

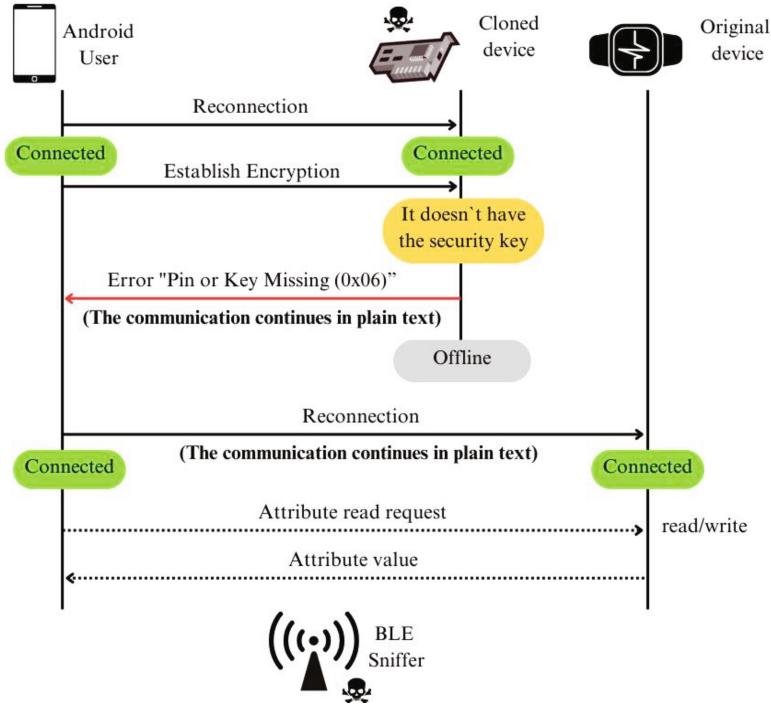


Fig. 3. Eavesdropping attack. After a reconnection with communication in plain text, a BLE Sniffer can capture data transmission.

According to the authors of [21] and [22], all the attacks mentioned so far can be carried out if there is no application-level security measure.

Relay Attack: Casagrande et al. [24] exploited the vulnerability during the COVID-19 pandemic to demonstrate that it is possible to perform a relay attack through government exposure-tracking applications, also known as Automatic Contact Tracing (ACT) applications. The authors claim that any ACT application using BLE is vulnerable to relay attacks by design: attackers can intercept and retransmit BLE signals involved in contact between two users. This opened the possibility for an attacker to cause a false positive for a victim by capturing exposure signals from an infected person and retransmitting them to their target.

Another vulnerability related to the BLE advertising process involves the freedom to set arbitrary values in the Service Data field of the Advertising Channel's Protocol Data Unit (PDU). PDUs are data packets used to transmit information between Bluetooth devices. These packets contain essential information for establishing connections, exchanging data, and controlling communication between devices (Fig. 6).

The authors of [23] exploited this vulnerability and designed an attack that allows extracting information from Android devices such as the operating system version, the latest security patch, contacts, calendar events, and even the list of installed applications. The main idea of the attack is to leverage the Service Data field to transport instructions.

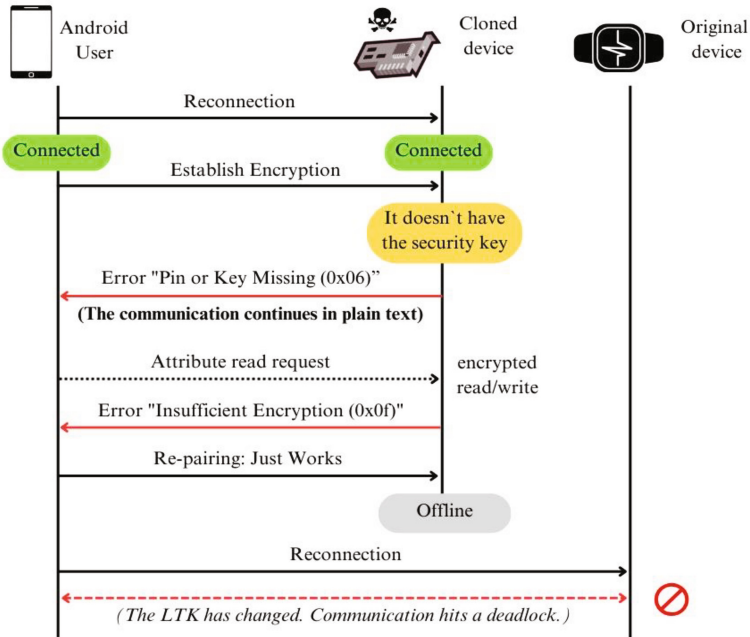


Fig. 4. DDoS attack. The LTK is changed by a cloned device (attacker), so original device cannot reconnect.

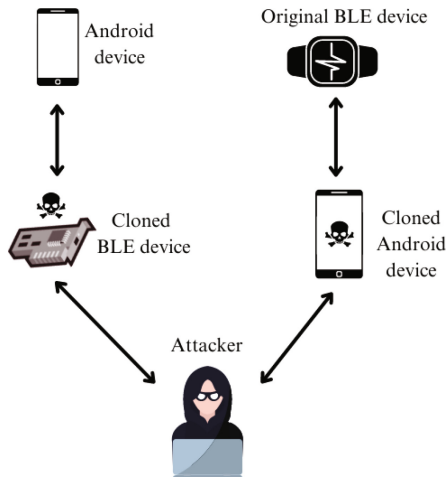


Fig. 5. Man-in-the-middle attack. An attacker cloned both devices to capture transmitted data.

The attacker advertises the type of service data with a fixed Service UUID (Universally Unique Identifier) and a Service Data field carrying commands encoded in bytes. On the other hand, the victim, through a malicious app installed on their Android device, scans BLE advertisements with the same Service UUID and receives the instructions. Due to

the nature of the attack, it can be carried out even if the devices are not paired and even if the victim’s device maintains a connection with other devices.

While vulnerabilities at the link level have been mentioned, there are also vulnerabilities at the application level. Google, on its official Android development site, warns that when a user links their device to another via BLE, the data communicated between them is accessible to all applications on the user’s device. Consequently, they recommend implementing application-level security measures [27].

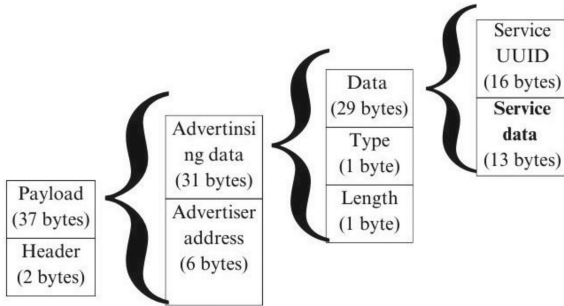


Fig. 6. Advertising channel PDU structure.

Despite this public warning, the authors of [20] conducted an analysis of a sample of 18,929 applications and found that 45% do not implement any type of application-level security. Most alarmingly, this rate increases to 70% for applications categorized as medical. The authors also found that, even if proper security measures are not taken, it may be possible to make changes to the device’s firmware.

In this way, Xiaomi implements application-level security measures for its devices. However, the authors of [26] have found six possible attacks with significant impact on its physical activity monitoring ecosystem. Four attacks allow wireless impersonation of any Xiaomi activity monitor and its corresponding application, conducting man-in-the-middle attacks and eavesdropping on their communications. The other two attacks use a malicious Android app to remotely eavesdrop on a monitor’s data, impersonating Xiaomi’s official physical activity app.

These attacks are possible due to the vulnerabilities that the authors expose in their work, among which the following stand out:

- Devices are not authenticated in the pairing process.
- The security key is shared in plain text.
- Only the application is authenticated in communication.
- Communication is not encrypted or authenticated.
- Communication is reliable.
- Weak user confirmation when pairing devices.

An important detail is that Xiaomi protects its physical activity monitoring ecosystem with custom application protocols communicated through BLE connections and TLS connections. However, it ignores the standard security mechanisms of BLE already supported by its devices.

Finally, another application-level vulnerability found by the authors of [19] allows any application to access the user's location while using BLE. However, they warn that Google is already aware of this vulnerability, and any application attempting to exploit it will be removed from the Play Store.

Table 4 summarizes the vulnerabilities and attacks found in the literature, establishing a correspondence between each attack and the vulnerabilities it exploits by using a cross matrix.

Table 4. Cross matrix of vulnerabilities and attacks found for BLE.

No.	Vulnerability	Level	Attacks							
			A1	A2	A3	A4	A5	A6	A7	
V1	The advertising device transmits the MAC address and attribute information in plaintext	Link	X	-	-	-	-	-	-	-
V2	The Service Data field of the Advertising Channel PDU can be leveraged to transmit custom messages	Link	-	-	X	-	-	-	-	-
V3	Weak user confirmation	Link	X	X	-	-	-	-	-	-
V4	Android does not provide the capabilities to detect and handle encryption errors	Link	X	X	-	X	X	-	-	-
V5	The Pairing Key is sent in plaintext	Application	X	X	-	-	-	-	-	-
V6	Unauthenticated pairing	Application	X	X	-	-	-	-	-	-
V7	Unilateral authentication in communication	Application	X	X	-	X	-	-	-	-
V8	Communication not encrypted or authenticated	Application	X	X	-	X	X	-	-	-
V9	Retransmittable communication	Application	X	-	-	X	-	X	X	-
V10	Ability to communicate with the peripheral device through co-located applications	Application	X	X	-	-	-	-	-	-

Note: A1: Impersonation Attacks. A2: Snooping Attacks. A3: Exfiltration Attacks. A4: Intermediary Attacks. A5: Denial of Service Attacks. A6: Replay Attacks. A7: Relay attacks

4.2 Proposal for Secure Communication for BLE

The vulnerabilities found in the systematic literature review were considered for development of the proposal. Although vulnerabilities were found at both the link and application levels, it was decided to propose a series of security measures at the application level. This choice was made because, at this layer, developers can implement customized and robust security measures. Our proposal primarily consists of encrypted communication and a semi-manual authentication process. After analyzing the selected works according to the methodology, it was concluded that these two features are essential to ensure secure BLE communication. Unlike link-level authentication, the proposed application-level authentication cannot be reduced to a non-interactive type similar to Just Works.

For the proposal, it is necessary to implement a service on the Generic Attribute (GATT) server, which is what BLE uses for data transmission. This is a client/server protocol in which the server has a collection of services the client can access. These services store data called characteristics. The service must have three characteristics: a public key of device “a” (PKa), a public key of device “b” (PKb), and an encryption key (SK), which will be used by the devices to establish secure communication.

The proposal consists of these security measures: blocking access to peripheral device data while not paired, manual authentication process, encrypted communication, communication with a key and timestamp, and renewal of encryption keys at random intervals. It is recommended to implement these security measures in 4 phases:

(1) Pre-pairing

This phase precedes the traditional pairing and runs each time the device is initiated (Fig. 7). Its purpose is to establish a security measure to prevent the peripheral device from transmitting its attributes in plaintext: if an attacker clones the Android device and connects to the peripheral device, they will not be able to read its attributes unless they impose their public key and establish secure communication with the peripheral device. However, this will only be possible if the attacker can access the hardware and pass the manual authentication process.

In this phase, the peripheral device waits to receive a public key in its corresponding field for encryption to initiate the subsequent phase.

(2) Authenticated Key Exchange

This phase is executed after the traditional pairing or whenever either of the devices has been formatted, and it can be considered an application-level pairing. The application verifies whether it has a stored encryption key during this process. If it does, it proceeds to the next phase. If not, it generates a pair of keys—public and private. It imposes its public key on the peripheral device, which reacts by generating its own pair of keys. This pair will be used to securely share the symmetric key and establish encrypted communication (Fig. 8).

The authors of [25] used the Lightweight Cryptography algorithm, ChaCha20, for BLE data transmission in their security monitoring proposal due to its suitability for dealing with limited bandwidth and energy.

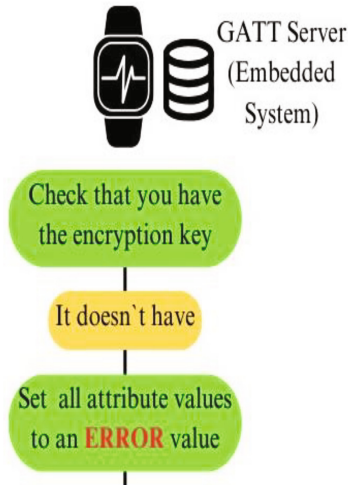


Fig. 7. Pre-pairing process. It prevents against an attack from a cloned android device.

(3) Communication

Once the devices have successfully shared the encryption key, they can communicate securely by encrypting each data they transmit. However, to ensure that the devices can recognize that the data is indeed encrypted and that they are using the same key, it is proposed to embed a key in each transmitted data. This key will be generated from the symmetric key (Fig. 8).

Various actions can be used to prevent replay attacks, including timestamps. This measure is recommended for its practicality and ease of implementation. Thus, they decrypt the content whenever the devices communicate and verify the key and timestamp (Fig. 9), sent in a structure like this:

$$\{\text{key}|\text{timestamp}|\text{content}\}$$

If the received data lacks a key or is different from the expected one, it is recommended that the Pre-pairing process be started again (Fig. 7). If this is not the case, the Authenticated Key Exchange process could be continued (Fig. 8). On the other hand, if the timestamp is not appropriate, it could indicate a replay attack.

(4) Renewal of Encryption Keys

This phase is executed at random intervals defined by the application. Unlike the Authenticated Key Exchange process, in this phase, to avoid interaction with the user and pair with fake devices, the new encryption key is encrypted and shared with the current key. Only if the device does not have the security key, the Authenticated Key Exchange process should be executed (Fig. 10).

Analysis of the Proposal

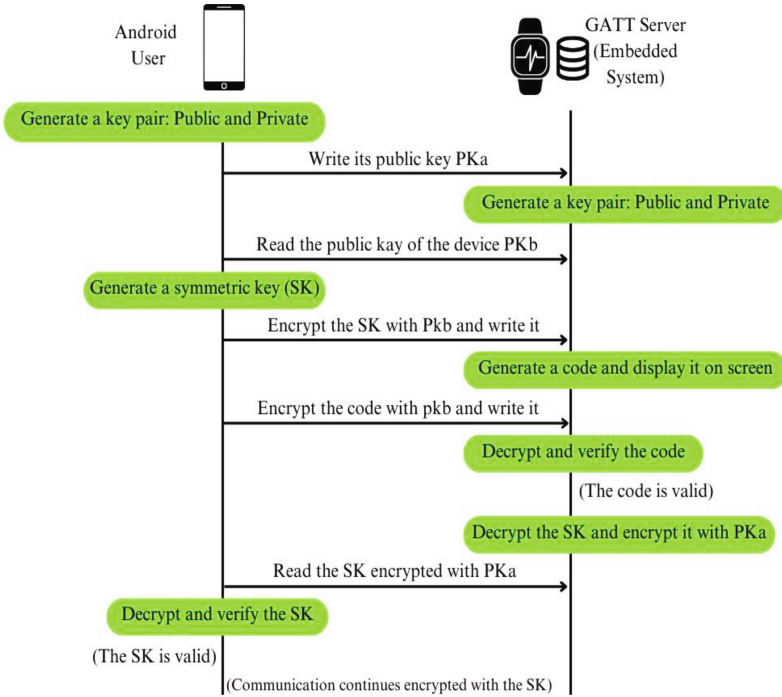


Fig. 8. Authenticated key exchange process. Both devices share their own public key, and transmit the symmetric key encrypted with the other device’s key.

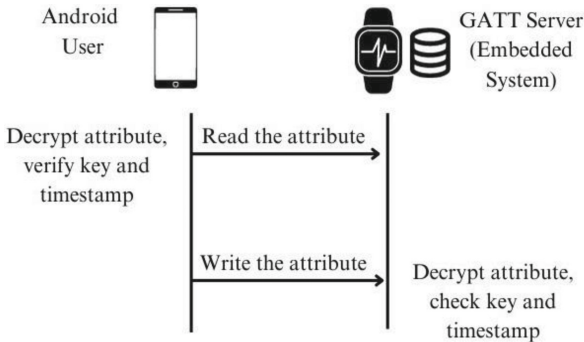


Fig. 9. Communication process. Devices use the symmetric key to communicate, and include a timestamp in the message.

The proposal was analyzed based on the attacks found in the literature review, and the possibility of attacks is briefly explained next:

- **Impersonation Attacks (A1):** Avoided since the transmitted data includes a key chosen when sharing the encryption key, thanks to manual authentication at the application level.

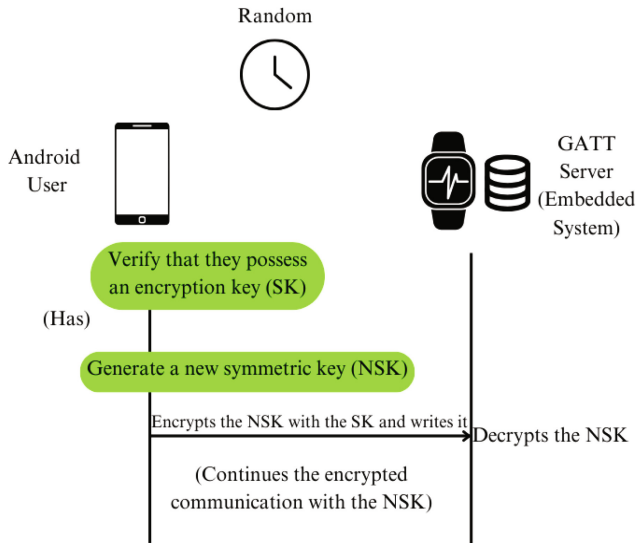


Fig. 10. Encryption key renewal process. The Android device generate a new symmetric key, encrypted it with current SK, and transmitted it to embedded device.

- **Snooping Attacks (A2):** Avoided since the data is transmitted encrypted.
- **Exfiltration Attacks (A3):** It is still possible to exfiltrate data from the operating system by exploiting the vulnerability at the link level concerning the PDU of the advertising channel. Until now, we have not found any application-level security measures to prevent this attack.
- **Intermediary Attacks (A4):** Avoided since the data is transmitted encrypted and due to manual authentication at the application level.
- **Denial of Service Attacks (A5):** It still is possible. As mentioned before, this attack exploits a series of vulnerabilities at the link layer; unfortunately, no solution was found.
- **Replay Attacks (A6):** Avoided since the data includes a timestamp.
- **Relay attacks (A7):** Avoided since the data includes a timestamp.

4.3 Discussion

The original Bluetooth specification aimed for connecting with devices without seriously considering the security of the transmitted data. Although security mechanisms have been added over time, they still do not entirely ensure communication. In the case of BLE, mainly when personal data is transmitted, security is a critical issue. Despite being a very widespread protocol, it is still impossible to solve all its security problems. Some commercial manufacturers are developing solutions for this issue but do not yet achieve an adequate security level because there are still vulnerabilities [26].

Some proposals can prevent some of the identified BLE attacks. For instance, the presented in [15] specifically prevent attacks A2, A4, A6 and A7 (four out of seven possible). However, that proposal is focused on BLE mesh networks and requires additional

infrastructure. On the other hand, the proposal of [16] could avoid attacks A2, A6 or A7, but it is unclear if it can be applied on Android devices.

The proposal presented in this paper is resilient against five out of the seven attacks (i.e. 71.43%) identified in the literature review, and it could be easily implemented in Android devices. Although the proposal is vulnerable to 28.57% of the attacks found, none of these affect the privacy or authenticity of the transmitted data. Therefore, the proposal fulfils its objective of ensuring secure communication.

5 Conclusions

BLE is a short-range, low-power wireless connection protocol increasingly used in IoT applications. To the date, several vulnerabilities have been identified in the protocol, which can put at risk user safety in the worst-case scenario. Many of these vulnerabilities currently have no solution. However, it is possible to leverage the application layer to implement customized security measures for mitigating the risks.

This work explored vulnerabilities and potential attacks on the protocol found in recent years, using them as a reference for designing a secure communication proposal that ensures privacy and data authenticity. The proposed solution can prevent five/of the seven attacks identified in the literature, accounting for 71.43% of the total. This scheme enhances the security, especially in applications handling critical or highly sensitive data.

As future work, the practical implementation of the protocol is pending to assess its effectiveness and impact on the performance of applications. Additionally, improvements may be possible to increase its security level or address the attacks it does not currently cover.

References





1. Sykes, E.R.: A context-aware system using mobile applications and beacons for on-premise security environments. *J. Ambient Intell. Human Comput.* **11**, 5487–5511 (2020). <https://doi.org/10.1007/s12652-020-01906-2>
2. Saikumar, P., Bharadwaja, P., Jabez, J.: Android and bluetooth low energy device based safety system. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1180–1185 (2019)
3. Apple, Google: Exposure Notification System, <https://www.apple.com/covid19/contacttracing>
4. Government of Singapore: BlueTrace, <https://bluetrace.io>
5. Bluetooth SIG, Inc.: 2022 Bluetooth Market Update, <https://www.bluetooth.com/2022-market-update/>
6. Soderi, S.: Cybersecurity Assessment of the Polar Bluetooth Low Energy Heart-Rate Sensor. In: Mucchi, L., Hämmäläinen, M., Jayousi, S., Morosi, S. (eds.) *Body area networks: Smart IoT and big data for intelligent health management*, pp. 252–265. Springer International Publishing, Cham (2019)
7. Kalinin, E., Belyakov, D., Bragin, D., Konev, A.: IoT security mechanisms in the example of BLE. *Computers* **10**, 162 (2021). <https://doi.org/10.3390/computers10120162>
8. StatCounter GlobalStats: Mobile Operating System Market Share Worldwide. <https://gs.statcounter.com/os-market-share/mobile/worldwide>

9. Căsar, M., Pawelke, T., Steffan, J., Terhorst, G.: A survey on bluetooth low energy security and privacy. *Comput. Netw.* **205**, 108712 (2022). <https://doi.org/10.1016/j.comnet.2021.108712>
10. Venkata Bhaskara Sastry, T., Amritha, P.P.: Bluetooth low energy devices: attacks and mitigations. In: Sengodan, T., Murugappan, M., Misra, S. (eds.) *International Conference on Advances in Electrical and Computer Technologies*, pp. 381–389. Springer, Singapore (2021)
11. Wang, J., Hu, F., Zhou, Y., Liu, Y., Zhang, H., Liu, Z.: BlueDoor: breaking the secure information flow via BLE vulnerability. In: *18th International Conference on Mobile Systems, Applications, and Services*, pp. 286–298. Association for Computing Machinery, New York, NY, USA (2020)
12. Pallavi, S., Narayanan, V.A.: An overview of practical attacks on BLE based IOT devices and their security. In: *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp. 694–698 (2019)
13. Kurt Peker, Y., Bello, G., Perez, A.J.: On the security of bluetooth low energy in two consumer wearable heart rate monitors/sensing devices. *Sensors* **22**, 988 (2022). <https://doi.org/10.3390/s22030988>
14. Klieme, E., Wilke, J., van Dornick, N., Meinel, C.: FIDOnuous: A FIDO2/WebAuthn extension to support continuous web authentication. In: *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1857–1867 (2020)
15. Banani, S., Thiemjarus, S., Wongthavarawat, K., Ounanong, N.: A dynamic light-weight symmetric encryption algorithm for secure data transmission via BLE beacons. *JSAN* **11**, 2 (2021). <https://doi.org/10.3390/jsan11010002>
16. Sun, D., Tian, Y.: Address privacy of bluetooth low energy. *Mathematics* **10**, 4346 (2022). <https://doi.org/10.3390/math10224346>
17. Yang, J., Poellabauer, C., Mitra, P., Neubecker, C.: Beyond beaconing: emerging applications and challenges of BLE. *Ad Hoc Netw.* **97**, 102015 (2020). <https://doi.org/10.1016/j.adhoc.2019.102015>
18. Bluetooth SIG, Inc.: Core Specification. <https://www.bluetooth.com/specifications/specs/core-specification-5-3/>
19. Toubiana, V., Cunche, M.: No need to ask the android: bluetooth-low-energy scanning without the location permission. In: *14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pp. 147–152. Association for Computing Machinery, New York, NY, USA (2021)
20. Sivakumaran, P., Blasco, J.: A Study of the feasibility of co-located app attacks against BLE and a large-scale analysis of the current application-layer security landscape. In: *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1–18 (2019)
21. Wu, J., Nan, Y., Kumar, V., Tian, D. (Jing), Bianchi, A., Payer, M., Xu, D.: BLESAs: spoofing attacks against reconnections in bluetooth low energy. In: *14th USENIX Workshop on Offensive Technologies (WOOT 20)* (2020)
22. Zhang, Y., Weng, J., Dey, R., Jin, Y., Lin, Z., Fu, X.: breaking secure pairing of bluetooth low energy using downgrade attacks. In: *29th USENIX Security Symposium (USENIX Security 20)*, pp. 37–54 (2020)
23. Gangwal, A., Singh, S., Spolaor, R., Srivastava, A.: BLEWhisperer: exploiting BLE advertisements for data exfiltration. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 13554 LNCS, pp. 698–717 (2022). https://doi.org/10.1007/978-3-031-17140-6_34
24. Casagrande, M., Conti, M., Losiouk, E.: Contact tracing made un-relay-able. In: *Eleventh ACM Conference on Data and Application Security and Privacy*, pp. 221–232. ACM, Virtual Event USA (2021)
25. Hernández-Goya, C., Aguasca-Colomo, R., Caballero-Gil, C.: BLE-based secure tracking system proposal. *Wireless Netw.* (2023). <https://doi.org/10.1007/s11276-023-03347-z>

26. Casagrande, M., Losiouk, E., Conti, M., Payer, M., Antonioli, D.: BreakMi: Reversing, exploiting and fixing xiaomi fitness tracking ecosystem. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pp. 330–366 (2022). <https://doi.org/10.46586/tches.v2022.i3.330-366>
27. Android Developers: Bluetooth Low Energy. <https://developer.android.com/guide/topics/connectivity/bluetooth/ble-overview>



Modeling and Control of a Peltier Thermoelectric System Applying a Multi-objective Optimization Approach

Víctor Huilcapi^(✉) , Geovanny García , Elias Ghia , and Brian Soto 

Universidad Politécnica Salesiana, Guayaquil, Ecuador
{vhuilcapi,ggarciaf1}@ups.edu.ec, {eghia,bsoto1}@est.ups.edu.ec

Abstract. This paper proposes a design procedure based on multi-objective optimization for modeling a Peltier thermoelectric system using experimental data. A multi-objective evolutionary algorithm was used to identify a set of optimal parameters that satisfactorily characterize the system's dynamics. The proposed methodology offers a designer valuable information about the dynamics of the temperatures of the Peltier's hot and cold surfaces and the trade-offs between their design objectives (visualized in the Pareto fronts). In this way, a control engineer can be sufficiently informed to choose, according to their preferences, a model for the Peltier cell with the best performance (for different system operation scenarios). A Peltier cold-side temperature model was selected to tune a proportional-integral-derivative (PID) controller and evaluate the robustness of the model. The tuned PID controller works directly on the nonlinear model of the Peltier cell, allowing it to effectively control the temperature of the cold surface of the thermoelectric module. The methodology uses the Integral Absolute Error (IAE) as a performance index to evaluate the quality of system modeling. The results show that the methodological approach applied to model and control the system performs very satisfactorily.

Keywords: Peltier cell · Multi-objective optimization · Pareto front · Multi-objective evolutionary algorithm · Nonlinear model · PID control

1 Introduction

Peltier cells are thermoelectric devices that use the Peltier-Seebeck effects to convert electrical energy into thermal energy and vice versa [1, 2]. By applying a voltage to a Peltier cell, it can generate a temperature gradient between its surfaces, that is, one surface of the Peltier cools while the other heats up. [3, 4]. Among the main advantages of Peltier cells we can mention that they do not use refrigerants or toxic gases polluting the environment to generate a temperature gradient (mainly the Peltier cell is required to cool) [5]. It also does not

Supported by Universidad Politécnica Salesiana.

require the use of compressors, condensers, or evaporators that generate noise and vibrations. Due to the multiple advantages that Peltier cells present, their applications have progressively increased over time. Among the most relevant applications are the cooling of microelectronic devices [6], refrigeration in medical applications [7], and refrigeration in the food industry [8].

This paper proposes a design method that applies a multi-objective optimization approach to model the dynamics of a Peltier thermoelectric cell and subsequently tune a PID controller to stabilize the temperature of a portable thermoelectric refrigerator. The modeling and effective control of the Peltier module are fundamental aspects for the optimal functioning of processes that use thermoelectric refrigeration [9, 10]. Thermoelectric refrigerators are currently very useful for transporting medications or vaccines sensitive to temperature changes. For example, during the COVID-19 pandemic, it was necessary to guarantee compliance with standards and regulations in the cold chain when carrying out portable transportation of the different types of vaccines created to mitigate the disease. A key aspect to guarantee the effectiveness of the vaccines was their transportation at appropriate temperatures [11, 12]. A relevant aspect for the effective use of thermoelectric cooling with Peltier cells is the creation of precise models that characterize the dynamics of these devices in different operating ranges. Engineers face a challenge in modeling Peltier thermoelectric devices due to the complexity of the thermal and electrical phenomena involved. Currently, different methods, such as machine learning and optimization algorithms, are used to generate advanced nonlinear models of Peltier thermoelectric devices [13, 14]. In the methodological approach proposed in this article, a multi-objective evolutionary algorithm (MOEA) is used in the optimization process to generate a robust Pareto front with models that satisfactorily characterize the dynamics of the Peltier system. Experimental data and performance indicators are also used to evaluate the quality and reliability of the models obtained for the Peltier system. The proposed methodology performs a multi-objective analysis of the system's variables, which provides valuable information to a designer for making well-informed decisions. In contrast, traditional optimization methods typically merge all of a system's information into a single design objective, which is likely to result in a loss of information relevant to understanding the system. The paper is structured as follows: Sect. 2 shows some fundamentals about multi-objective optimization. Section 3 describes the Peltier system and the nonlinear model that characterizes its dynamics. Section 4 shows the methodology applied for the identification of nonlinear models of the Peltier system and temperature control. Section 5 corresponds to the analysis of the results obtained in the multi-objective identification process of Peltier system models. Finally, the conclusions are shown in Sect. 6.

2 Multi-Objective Optimization Fundamentals

A design procedure with a multi-objective optimization (MOOD) approach is made up of at least three stages: 1) the definition of the multi-objective problem

(MOP), 2) the multi-objective optimization process, and 3) the multi-criteria decision stage [15, 16]. In the first stage, MOP can be defined as shown in (1)–(5).

$$\min_{\mathbf{x} \in D} \mathbf{J}(\mathbf{x}, \boldsymbol{\psi}) \tag{1}$$

$$\mathbf{J}(\mathbf{x}, \boldsymbol{\psi}) = \{J_1(\mathbf{x}, \boldsymbol{\psi}), J_2(\mathbf{x}, \boldsymbol{\psi}), \dots, J_k(\mathbf{x}, \boldsymbol{\psi})\} \tag{2}$$

$$G(\mathbf{x}) \leq 0 \tag{3}$$

$$H(\mathbf{x}) = 0 \tag{4}$$

$$\underline{\mathbf{x}} \leq \mathbf{x}_i \leq \bar{\mathbf{x}}, i = [1, \dots, n] \tag{5}$$

where, $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ is the decision vector; D is the decision space; $\mathbf{J}(\mathbf{x}, \boldsymbol{\psi}) \in \mathbf{R}^m$ is the design objectives vector; $\boldsymbol{\psi}$ represents possible uncertainties in the system, $G(\mathbf{x}), H(\mathbf{x})$ are the MOP restrictions; $\underline{\mathbf{x}}, \bar{\mathbf{y}}$ are the lower and upper bounds of the search space. In the second stage, a multi-objective optimization algorithm must be chosen to satisfactorily obtain the approximations of the Pareto fronts of each MOP. For this work, different optimization algorithms have been developed over time, as shown in [17–20]. Finally, in the third stage, it is necessary to analyze the characteristics or trade-offs of the solutions that are part of the Pareto front $\mathbf{J}(\mathbf{X}_p^*)$ to choose the best solution according to the designer’s preferences (since all the solutions of the Pareto front are optimal). There are various multidimensional Pareto front visualization tools for this task, as shown in [21–23].

3 System Description

A Peltier cell is a thermoelectric device with two surfaces, as shown in Fig. 1. Applying a DC voltage to the Peltier cell cools one surface, and the other is heated. A Peltier cell uses the *Seebeck* effect to convert heat into electrical current and the *Peltier* effect to convert electrical current into a temperature gradient [24, 25]. An energy balance on each surface of the Peltier and the application of the first law of thermodynamics allow us to obtain a simple nonlinear model that characterizes the dynamics of the temperatures of the cold and hot sides of the system. The Peltier cell model is shown in (6)–(13).

The energy balance on the cold surface of the Peltier is shown in (6).

$$Q_{ncs} = Q_{ecs} - Q_{acs} - Q_{jou} + Q_{hcs} \tag{6}$$

Q_{ncs} represents the net heat absorbed by the cold side of the Peltier; Q_{ecs} is the heat transmitted by convection between the environment and the cold face; Q_{acs} is the heat absorbed by the cold face due to the Peltier effect; Q_{jou} is the heat generated by the Peltier due to the Joule effect; Q_{hcs} is the heat transmitted by conduction from the hot face to the cold face.

$$Q_{ncs} = mce_{cool} \frac{d}{dt} T_{cool}(t) \tag{7}$$

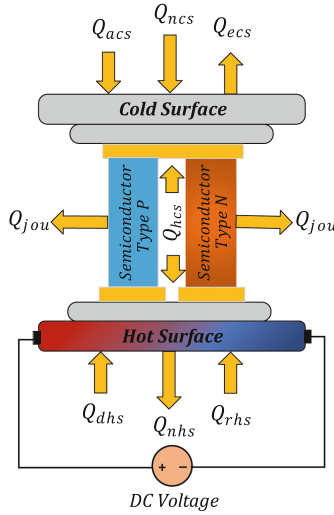


Fig. 1. Block diagram and energy flow diagram of the Peltier cell.

$$T_{cool}(t) = \frac{1}{mce_{cool}} \int_{t_o}^{t_f} \left[(T_{amb} - T_{cool})K_{conva} - \alpha T_{cool}I_p - \frac{1}{2}I_p^2 R_p + (T_{hot} - T_{cool})K_{convp} \right] dt \tag{8}$$

The energy balance on the hot surface of the Peltier is shown in (9).

$$Q_{nhs} = Q_{rhs} + Q_{dhs} + Q_{jou} - Q_{hcs} \tag{9}$$

Q_{nhs} is the net heat dissipated by the hot side of the Peltier; Q_{rhs} is the heat transmitted between the radiator and the hot side; Q_{dhs} is the heat dissipated on the hot side due to the Peltier effect; Q_{hcs} represents the net heat flow in the radiator; Q_{cer} is the heat transmitted by convection between the environment and the radiator. All these variables are measured in Watt.

$$Q_{nhs} = mce_{hot} \frac{d}{dt} T_{hot}(t) \tag{10}$$

$$T_{hot}(t) = \frac{1}{mce_{hot}} \int_{t_o}^{t_f} \left[(T_{rad} - T_{hot})K_{rad} + \alpha T_{hot}I_p + \frac{1}{2}I_p^2 R_p - (T_{hot} - T_{cool})K_{convp} \right] dt \tag{11}$$

The energy balance between the hot surface of the Peltier and the attached radiator is shown in (12).

$$Q_{hsr} = Q_{cer} - Q_{rhs} \tag{12}$$

$$Q_{hsr} = mce_{rad} \frac{d}{dt} T_{rad}(t) \quad (13)$$

The physical representation of each parameter of the proposed model for the Peltier cell is shown in Table 1.

Table 1. Description of Peltier cell model parameters.

Parameter	Parameter description	Units
K_{fan}	Factor of convection between the environment and the radiator	[W/°K]
T_{amb}	Environment temperature	[°C]
T_{cool}	Temperature on the cold surface of the Peltier	[°C]
T_{hot}	Temperature on the hot surface of the Peltier	[°C]
T_{rad}	Temperature in the radiator coupled to the Peltier	[°C]
α	Seebeck coefficient	[V/°K]
I_p	Current circulating in the Peltier cell	[A]
R_p	Peltier cell resistance	[Ω]
K_{convp}	Thermal coefficient of the Peltier	[W/°K]
K_{rad}	Thermal coefficient between the hot face and the radiator	[W/°K]
K_{conva}	Thermal coefficient between the hot face and the environment	[W/°K]
mce_{cool}	Thermal capacity in the cold face	[J/°K]
mce_{hot}	Thermal capacity in the hot face	[J/°K]
mce_{rad}	Thermal capacity in the radiator	[J/°K]

The experimental system is shown in Fig. 2. It is mainly composed of a thermoelectric refrigerator prototype, inside which is the Peltier cell, and a programmable controller PLC S7-1200, that regulates the temperatures of both sides of the Peltier through a controllable voltage source. Data acquisition and system control are done with an Intel Core I7 Processor using MATLAB/SIMULINK and TIAPortal tools. The operation of the prototype is shown in the following link: <https://www.youtube.com/watch?v=srLqm7EgjHw>.

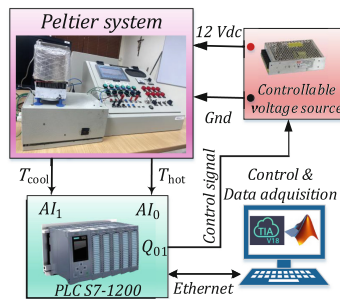


Fig. 2. Experimental Peltier system.

4 Methodology

A multi-objective optimization problem (MOP) to identify a set of optimal Peltier cell models is posed in (14)–(19). To propose the MOP, the parameters to be identified from the Peltier model must be considered, as well as the system’s restrictions or operating ranges. An optimization criteria must also be taken into consideration, in this case, the average error of the temperatures on the cold and hot sides of the Peltier system was used. Laboratory tests were performed for 2100 seconds (in the model identification experiment) and 1000 s (in the model validation experiment), and 10500 samples ($N = 10500$) were acquired.

$$\min_{\mathbf{x}_m \in D} \mathbf{J}(\mathbf{x}_m) \tag{14}$$

$$\mathbf{J}(\mathbf{x}_m) = \{J_1(\mathbf{x}_m), J_2(\mathbf{x}_m)\} \tag{15}$$

$$J_1(\mathbf{x}_m) = \frac{1}{N} \sum_{k=1}^N |T_{cold.experimental}(k) - T_{cold.identified}(k)| \tag{16}$$

$$J_2(\mathbf{x}_m) = \frac{1}{N} \sum_{k=1}^N |T_{hot.experimental}(k) - T_{hot.identified}(k)| \tag{17}$$

The decision vector is represented by:

$$\mathbf{x}_m = \{\alpha, R_p, K_{convp}, K_{rad}, K_{conva}, mce_{cool}, mce_{rad}, mce_{hot}\} \tag{18}$$

$$\underline{\mathbf{x}}_m \leq \mathbf{x}_m \leq \bar{\mathbf{x}}_m \tag{19}$$

where $\underline{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_m$ were established taking into consideration the technical information provided by the manufacturer of the Peltier system, (see Table 2).

Table 2. Bounds of the decision vector \mathbf{x}_m for the identification of Peltier models.

	Bounds of \mathbf{x}_m							
\mathbf{x}_m	α	R_p	K_{convp}	K_{rad}	K_{conva}	mce_{cool}	mce_{rad}	mce_{hot}
$\underline{\mathbf{x}}_m$	0.01	0.5	0.01	0.1	0.01	0.1	10	0.1
$\bar{\mathbf{x}}_m$	0.5	1.5	50	5	5	20	10000	50

A Proportional-Integral-Derivative (PID) control structure was used to control the temperature of the Peltier’s cold surface, as shown in (20).

$$u(t) = K_p e(t) + \frac{K_p}{T_i} \int_0^t e(t) dt + K_p T_d \frac{de(t)}{dt} \tag{20}$$

where K_p is the proportional gain, T_i is the integral time, T_d is the derivative time, and $e(t)$ is the tracking error between the output and the input.

The multi-objective evolutionary algorithm *ev – MOGA* was used in the optimization stage to identify the set of nonlinear models of the Peltier cell [19, 20]. The *ev – MOGA* algorithm is an elitist multi-objective evolutionary algorithm based on the concept of ϵ -dominance, where two subpopulations evolve simultaneously and iteratively to effectively achieve convergence to the Pareto front. The configuration of the *ev – MOGA* is shown in algorithm 1.

Algorithm 1 *ev – MOGA* algorithm configuration

```

eMOGA.objfun = fcoste multi – objective Peltier
eMOGA.objfun.dim = 3
eMOGA.search space UB = [0.1 1.5 1.5 10 1 50 1000 15]
eMOGA.search space LB = [0.01 0.1 0.01 0.1 0.1 0.1 0.1 0.1]
eMOGA.Generations = 1000
eMOGA.n.div = 100
[pfront, pset, eMOGA] = evMOGA(eMOGA)
function : J = fcoste multi – objective Peltier(model, param)
alpha = model(1); Rpeltier = model(2); Kconvp = model(3); Krad = model(4)
Kconva = model(5); mcecool = model(6); mcerad = model(7); mcehot = model(8)
simulink (peltiermodel_experimental)
J1 = J1(xm)
J2 = J2(xm)
J = [J1 J2]

```

In the visualization stage of Pareto fronts and decision-making, the *Level Diagram* (LD) tool was used [23]. LD is based on classifying the Pareto front $\mathbf{J}(\mathbf{X}_p)$ according to its proximity to the ideal point. The ideal solution is understood as the point with minimum values on the Pareto front for each design objective of a MOP. The LD tool evaluates the distance of a solution from the Pareto front to the ideal point using a given norm (*1 – norm*, *2 – norm*, *∞ – norm*). Each norm offers the designer different diagram characteristics and provides various points of view. A classical way of representing the Pareto front of a two-dimensional MOP is shown in Fig. 3, and the same Pareto front is shown in Fig. 4 using the LD tool with *1 – norm*. The LD tool assigns an independent graph to each objective, J_1 and J_2 . It does so in an equivalent way with each MOP decision variable. The y-axis of each graph corresponds to the value of the norm selected to represent the Pareto front, and all graphs are synchronized concerning this axis. The x-axis corresponds to the values of the MOP design objectives (in the objectives space) or the value of the decision variables (in the decision space).

5 Results and Discussion

5.1 Results of Multi-objective Identification of Peltier Cell Models

Laboratory tests were carried out to obtain a set of models of the Peltier cell. The experimental data acquired were used to identify a set of nonlinear models for the

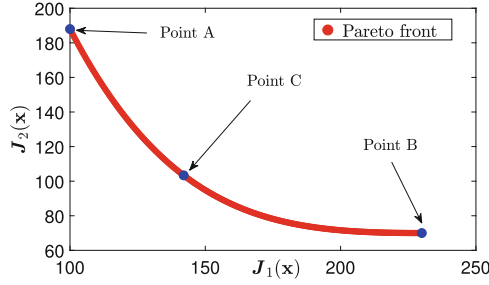


Fig. 3. Classic representation of a two-dimensional Pareto front.

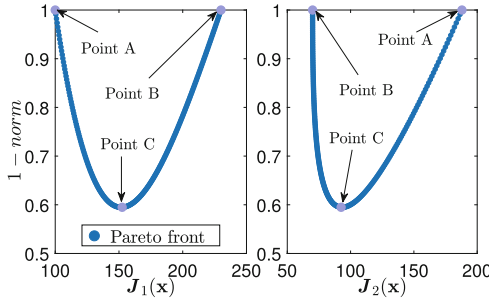


Fig. 4. Two-dimensional Pareto front using LD with 1-norm.

temperatures of the hot and cold faces. The Pareto fronts obtained in the multi-objective optimization process are shown in Fig. 5. Each symbol (pink, blue and yellow solutions) in the Pareto fronts represents a model of the Peltier system with specific characteristics. Likewise, Fig. 6 shows the identified parameters of each model obtained for the Peltier system (the parameters are described in Table 3).

Analyzing the Pareto fronts in Fig. 5, it is possible to observe a conflict between the models of the Peltier hot and cold side temperatures since the pink and blue stripes are on opposite sides of each Pareto front. In other words, the pink solutions represent the best models to characterize the cold side temperature, but the worst models for the hot side temperature. Likewise, the yellow solutions represent the best models for the Peltier hot side temperature but the worst models for characterizing the cold side.

Table 3. Model parameters selected to characterize the Peltier.

α	R_p	K_{convp}	K_{rad}	K_{conva}	mce_{cool}	mce_{rad}	mce_{hot}
0.04	0.82	0.19	9.59	0.5	9.25	730	13.02

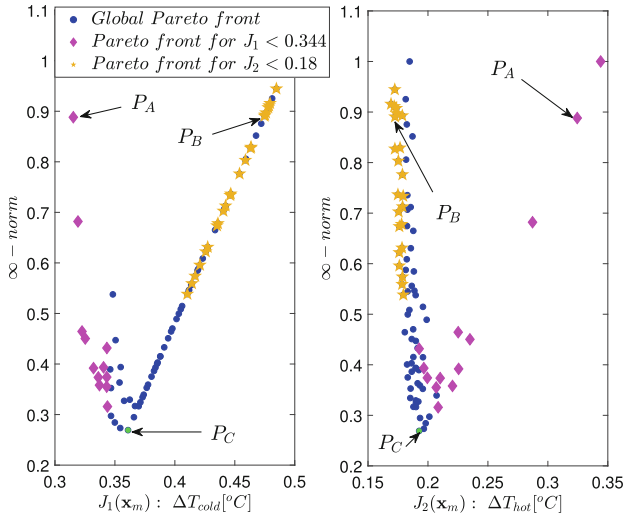


Fig. 5. Pareto front of the set of models identified for the Peltier system.

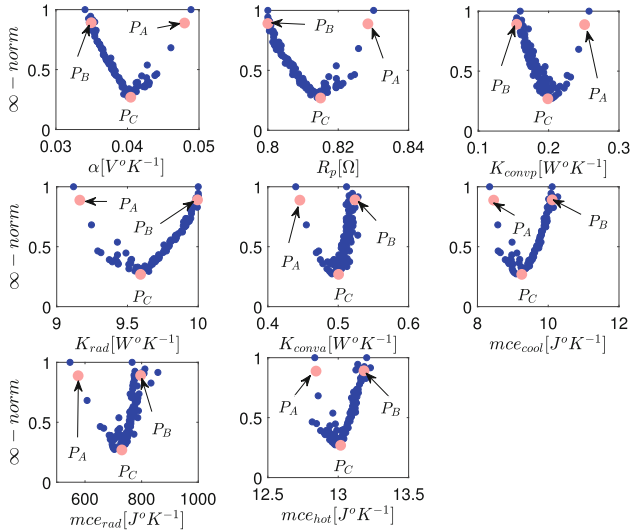


Fig. 6. Identified parameters of the nonlinear model of the Peltier system. Each graph represents a Pareto-optimal parameter of the Peltier model in a range of values established in the multi-objective optimization process (according to technical characteristics). A designer can select a model for the system based on preferences or the operating scenario (for example P_A , P_B , P_C).

The validation of the temperatures of the cold and hot sides of the Peltier system on the identification data is shown in Fig. 7. It can be seen that the set of models obtained (88 models) in the optimization process satisfactorily fit the real dynamics of the system. To ensure the reliability of the models obtained for the Peltier, they were validated on a different section of data (the data for model validation are different from those used for identification.), as shown in Fig. 8. It is observed that even when varying the dynamics of the system, the models obtained (in the process of multi-objective identification) satisfactorily represent the system temperatures (although, logically, they have a greater approximation error). The statistics of the errors in the models of the temperatures of the cold and hot sides of the Peltier system are shown in Fig. 9 and Table 4. As can be seen, the models found to represent the Peltier system have an average error less than 0.6 °C. Also, 75% of the models have an error of less than 0.7 °C, and all models have an error of less than 1 °C.

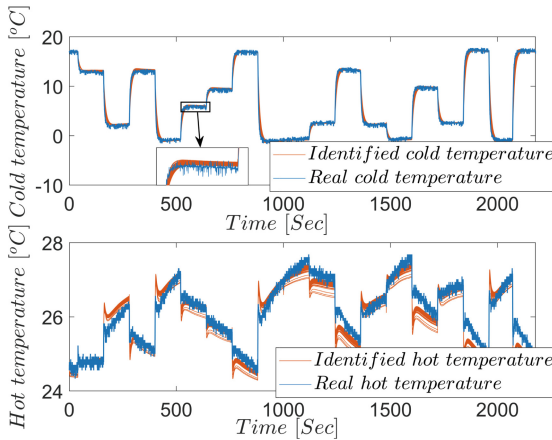


Fig. 7. Models identified for the cold and hot temperatures of the Peltier system.

Table 4. Errors in the cold and hot temperature models of the Peltier system.

Peltier temperature	E_{min}	E_{max}	E_{mean}	E_{median}
$T_{cold-identification\ data}$	0.31	0.49	0.40	0.39
$T_{cold-validation\ data}$	0.43	0.53	0.47	0.46
$T_{hot-identification\ data}$	0.17	0.34	0.19	0.18
$T_{hot-validation\ data}$	0.37	0.97	0.59	0.60

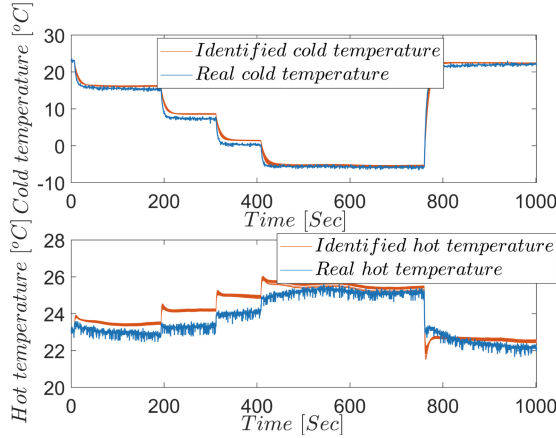


Fig. 8. Models validated for the cold and hot temperatures of the Peltier system.

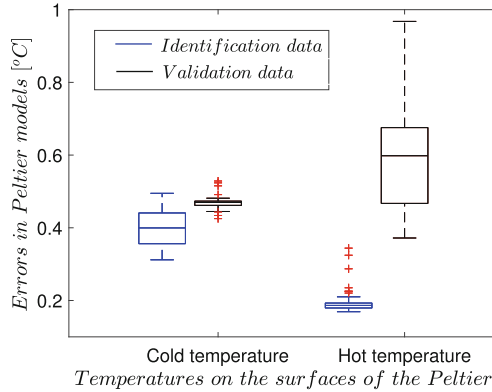


Fig. 9. Box plot of Peltier model errors. The blue boxes show the errors in the Peltier models on the same identification data set, and the black boxes show the errors of the models validated in another system operation scenario.

To evaluate in detail the trade-offs and performances between Peltier models belonging to each of these regions (pink, blue, and yellow regions), three models (models in P_A , P_B , and P_C) were selected on the Pareto fronts, and their performances are shown in Table 5. As shown in Table 5, the model in P_A is the best to characterize the cold face temperature since $J_1(\mathbf{x}_m)_{P_A} < J_1(\mathbf{x}_m)_{P_B}$ and $J_1(\mathbf{x}_m)_{P_A} < J_1(\mathbf{x}_m)_{P_C}$, but it is the worst to represent the hot face temperature since $J_2(\mathbf{x}_m)_{P_A} > J_2(\mathbf{x}_m)_{P_B}$ and $J_2(\mathbf{x}_m)_{P_A} > J_2(\mathbf{x}_m)_{P_C}$. Similarly, the model in P_B is the best for characterizing the Peltier hot face temperature since $J_2(\mathbf{x}_m)_{P_B} < J_2(\mathbf{x}_m)_{P_A}$ and $J_2(\mathbf{x}_m)_{P_B} < J_2(\mathbf{x}_m)_{P_C}$ but the worst for representing the cold face temperature since $J_1(\mathbf{x}_m)_{P_B} > J_1(\mathbf{x}_m)_{P_A}$ and $J_1(\mathbf{x}_m)_{P_B} > J_1(\mathbf{x}_m)_{P_C}$. So also, if a designer wishes to select, according to his preferences,

a model for the temperatures of the cold and hot sides of the Peltier that are not the best but have satisfactory performances, the designer could select any other solution from the Pareto front. For example, the model in PC whose performance is in the ranges $J_1(\mathbf{x}_m)_{P_A} < J_1(\mathbf{x}_m)_{P_C} < J_1(\mathbf{x}_m)_{P_B}$ and $J_2(\mathbf{x}_m)_{P_B} < J_2(\mathbf{x}_m)_{P_C} < J_2(\mathbf{x}_m)_{P_A}$ was chosen on the Pareto front. The validation of the models in P_A , P_B and P_C are shown in Figs. 10 and 11 respectively.

Table 5. Performance of Peltier models in P_A, P_B, P_C selected on the Pareto front shown in Fig. 5.

$\mathbf{J}(\mathbf{x}_m)$	Model P_A	Model P_B	Model P_C
$J_1(\mathbf{x}_m)$	0.315	0.475	0.361
$J_2(\mathbf{x}_m)$	0.325	0.173	0.193

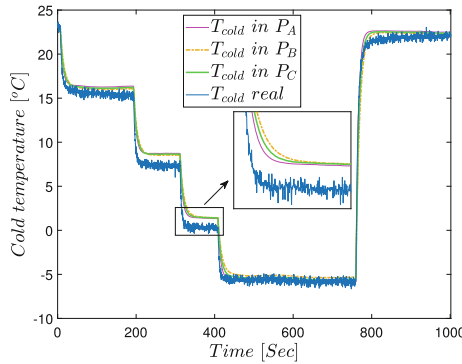


Fig. 10. Peltier cold side temperature models for points P_A, P_B , and P_C .

5.2 Peltier System Controller Tuning Details

Finally, the model in PC was selected to represent the dynamics of the Peltier cold and hot surface temperatures, and its parameters are shown in Table 3. Using the model in PC and *MATLAB-SIMULINK*, a PID controller was tuned for the Peltier cold face temperature. The tuning parameters were loaded into the PID compact of an *PLC* : *S7-1200* and are, $K_p = 60.91$, $T_i = 32.77$ [s] and $T_d = 8.29$ [s]. The Peltier cold face temperature control is shown in Fig. 12. The system works in different operating ranges to evaluate the performance of the controller and the effectiveness of the Peltier model. A satisfactory performance of the Peltier system is observed since the cold face temperature stabilizes in about 2 [min] and has no overshoot.

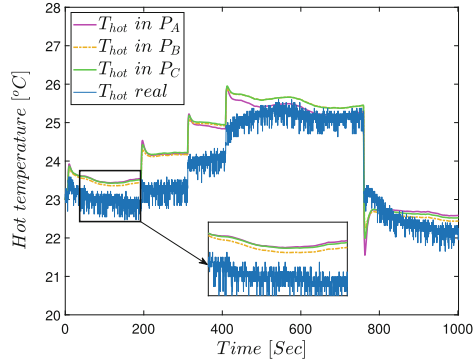


Fig. 11. Peltier hot side temperature models for P_A , P_B , and P_C .

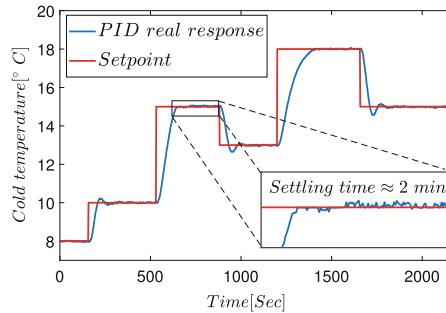


Fig. 12. Control of the cold temperature of the Peltier cell.

6 Conclusions

A methodology for modeling a Peltier cell applying a multi-objective optimization approach has been shown in this paper. Experimental results show that the set of nonlinear models obtained has a satisfactory performance since the error between experimental and simulated data is minimal (according to the statistics shown). The proposed methodological approach shows relevant information for a designer (about the performance and trade-offs between the variables of the Peltier models). In this way, a control engineer can select in an informed way the best model to represent the Peltier. For example, The methodology reveals a conflict between the models of the cold and hot surfaces of the Peltier since trying to improve one makes the other one worse (this information would be hidden in a traditional mono-objective approach). Finally, a compromise model was selected to represent the dynamics of the Peltier cell and to be able to tune a PID control for the cold side temperature. According to the results obtained, the PID control works satisfactorily since it stabilizes the system in a reasonable time and does not present an overshoot. When replicating this approach in another thermoelectric cooling prototype, it is recommended that sufficient system dynamics

data be acquired so that the multi-objective optimization algorithm can find a robust Pareto front (robust system models) and adjust controllers that effectively stabilize the system. The approach, while presenting the limitation of a high computational cost compared to traditional methods for modeling and designing controllers, it also offers a wealth of valuable information that is crucial for a designer's decision-making.

Future work will focus on building a more robust portable thermoelectric refrigerator, embedding the model and controller in hardware that provides autonomy to the prototype. Electrical autonomy for the new prototype is also planned for medical use in different geographical areas.

Acknowledgements. The authors thank to the Research Group In Control Systems and Robotics (GISCOR). This work has been supported by Universidad Politécnica Salesiana through project No. 0138-007-2023-12-12.



References

1. Freire, L.O., Navarrete, L.M., Corrales, B.P., Castillo, J.N.: Efficiency in thermoelectric generators based on Peltier cells. *Energy Rep.* **7**, 355–361 (2021)
2. Fallmann, M., Kölbl, J., Ausweger, T., Lösch, M., Poks, A., Kozek, M.: Test bed emulation of secondary loop refrigeration units using Peltier elements: an impedance control approach. *IFAC-PapersOnLine* **56**(2), 6435–6440 (2023)
3. Shi, L., Abed, A.M., Fayed, M., Abdulghani, Z.R., Anqi, A.E., Khadimallah, M.A., Moria, H., Wae-hayee, M.: Economic cost analysis of air-cooling process using different numbers of Peltier modules; experimental case study. *Case Stud. Thermal Eng.* **41**, 102627 (2023)
4. Abbas, Z., Shah, A., Hassan, M.T., Ali, M.S., Ud Din, Q., Naseem, B., Asghar, A., Haider, A.: Performance evaluation of novel solar-powered domestic air cooler with Peltier modules. *J. Mech. Sci. Technol.* **34**(11), 4797–4807 (2020)
5. Guélou, G., Couder, C., Manière, C., Candolfi, C., Lenoir, B., Lallemand, L., Coureau, C., Guilmeau, E.: Issues and opportunities from Peltier effect in functionally-graded Colusites: From SPS temperature modeling to enhanced thermoelectric performances. *Appl. Mater. Today* **22**, 100948 (2021)
6. Chen, Z.G., Liu, W.D.: Thermoelectric coolers: infinite potentials for finite localized microchip cooling. *J. Mater. Sci. Technol.* **121**, 256–262 (2022)
7. Zaferani, S.H., Sams, M.W., Ghomashchi, R., Chen, Z.G.: Thermoelectric coolers as thermal management systems for medical applications: design, optimization, and advancement. *Nano Energy* **90**, 106572 (2021)
8. Siddique, A.R.M., Bozorgi, M., Venkateshwar, K., Tasnim, S., Mahmud, S.: Phase change material-enhanced solid-state thermoelectric cooling technology for food refrigeration and storage applications. *J. Energy Storage* **60**, 106569 (2023)
9. Khan, A., Yaqub, S., Ali, M., Ahmad, A.W., Nazir, H., Khalid, H.A., Iqbal, N., Said, Z., Sopian, K.: A state-of-the-art review on heating and cooling of lithium-ion batteries for electric vehicles. *J. Energy Storage* **76**, 109852 (2024)
10. Ali, H.M., Rehman, T.u., Arıcı, M., Said, Z., Duraković, B., Mohammed, H.I., Kumar, R., Rathod, M.K., Buyukdagli, O., Teggat, M.: Advances in thermal energy storage: fundamentals and applications. *Progr. Energy Combustion Sci.* **100**, 101109 (2024)

11. Thakur, V., Hossain, M.K., Mangla, S.K.: Factors to vaccine cold chain management for sustainable and resilient healthcare delivery. *J. Clean. Prod.* **434**, 140116 (2024)
12. Cattin, M., Jonnalagedda, S., Makohliso, S., Schönenberger, K.: The status of refrigeration solutions for last mile vaccine delivery in low-income settings. *Vaccine: X* **11**, 100184 (2022)
13. Tiryaki, H., Yusuf, A., Ballikaya, S.: Determination of electrical and thermal conductivities of n-and p-type thermoelectric materials by prediction iteration machine learning method. *Energy*, 130597 (2024)
14. Ge, Y., He, K., Xiao, L., Yuan, W., Huang, S.M.: Geometric optimization for the thermoelectric generator with variable cross-section legs by coupling finite element method and optimization algorithm. *Renew. Energy* **183**, 294–303 (2022)
15. Huilcapi, V., Blasco, X., Herrero, J.M., Reynoso-Meza, G.: A loop pairing method for non-linear multivariable control systems under a multi-objective optimization approach. *IEEE Access* **8**, 41262–41281 (2020)
16. Huilcapi, V., Castillo, C., Sanchez, D., Cajo, R.: Optimization of an air pressure system: A multi-objective control and modeling approach. *IEEE Access* **12**, 96691–96703 (2024)
17. Wang, L., Yao, Y., Luo, X., Adenutsi, C.D., Zhao, G., Lai, F.: A critical review on intelligent optimization algorithms and surrogate models for conventional and unconventional reservoir production optimization. *Fuel* **350**, 128826 (2023)
18. Blekos, K., Brand, D., Ceschini, A., Chou, C.H., Li, R.H., Pandya, K., Summer, A.: A review on quantum approximate optimization algorithm and its variants. *Phys. Rep.* **1068**, 1–66 (2024)
19. Huilcapi, V., Blasco, X., Herrero, J.M., Pajares, A., Reynoso-Meza, G.: Application of an input-output pairings selection methodology to control multivariable systems based on multi-objective optimization. In: 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 443–449. *IEEE* (2021)
20. Herrero, J., Blasco, X., Martínez, M., Ramos, C., Sanchis, J.: Non-linear robust identification of a greenhouse model using multi-objective evolutionary algorithms. *Biosys. Eng.* **98**(3), 335–346 (2007)
21. Nagar, D., Ramu, P., Deb, K.: Visualization and analysis of pareto-optimal fronts using interpretable self-organizing map (ISOM). *Swarm Evol. Comput.* **76**, 101202 (2023)
22. Xu, Y., Zhang, H., Huang, L., Qu, R., Nojima, Y.: A pareto front grid guided multi-objective evolutionary algorithm. *Appl. Soft Comput.* **136**, 110095 (2023)
23. Blasco, X., Herrero, J.M., Reynoso-Meza, G., Iranzo, M.A.M.: Interactive tool for analyzing multiobjective optimization results with level diagrams. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 1689–1696 (2017)
24. Huilcapi, V., Herrero, J.M., Blasco, X., Martínez-Iranzo, M.: Non-linear identification of a Peltier cell model using evolutionary multi-objective optimization. *IFAC-PapersOnLine* **50**(1), 4448–4453 (2017)
25. Wang, L., Shu, L., Lv, Y., Hu, Q., Ma, L., Jiang, X.: Ultra-efficient and thermally controlled atmospheric structure deicing strategy based on the Peltier effect. *Appl. Thermal Eng.*, 123162 (2024)



Wearable Device for Acquiring Biomechanical Variables Applied to the Analysis of Occupational Health Risks in Industrial Environments

Carlos Calderon-Cordova^(✉) , Victor Puchaicela, and Roger Sarango 

Universidad Tecnica Particular de Loja, Loja 1101608, Ecuador
{cacalderon, vfpuchaicela, rasarango1}@utpl.edu.ec

Abstract. The activities performed in an industrial environment are classified as high risk for causing occupational diseases because the activities performed demand high frequencies of execution of movements or require high physical efforts on the part of industrial operators. Currently, risk indexes are assessed only based on observation and questionnaires, which adds subjectivity and lack of reliability to the results. The present study proposes the development and evaluation of a wearable technological device to acquire in situ the biomechanical variables of an operator and that these data serve to determine occupational risk indexes in a more objective way. The hardware architecture of the wearable device implemented consists of four modules: signal acquisition, processing, storage, and data transmission. The eleven (11) monitored variables are six myoelectric signals in the forearm, biceps, and back, two back inclination angles, heart rate, skin electrical conductance signal, and body temperature. To evaluate the performance of the implemented device, the eleven monitored variables were acquired, registered, and visualized during the execution of three test exercises by three users of different physiological parameters, resulting in a set of 99 signals. It was observed congruence between the occurrence of peaks of the signals with the characteristics of the executed exercises, in addition, there is no loss of data between the stages of capture, transmission, recording, and visualization.

Keywords: Wearables · biomechanical signals · occupational health · Internet of Medical Things (IoMT) · embedded systems

1 Introduction and Related Work

The activities realized in an industrial environment are classified as high risk, because the workers are in continuous contact with machinery and mechanical equipment or because they perform repetitive activities that are hazardous to their health or require high physical effort. Due to these activities that demand high frequencies of execution or that require high physical efforts by industrial operators, there is a high risk of causing occupational diseases. The density of occurrence of occupational diseases can be considered a public health problem and is also considered an economic problem for the industrial production sector of a country [1, 2].

To mitigate the consequences of occupational diseases, the industrial production sector determines risk indexes for each activity performed by workers. However, these risk indexes are determined on the basis of visual verifications and questionnaires, which leads to subjective and unreliable results [3, 4]. On the other side, there are assessments based on medical instrumentation, however, this is considered invasive for the required purpose, since it completely modifies the work environment because the operator under analysis must emulate his activity within a laboratory environment [5–7].

Based on the previous analysis, the present project proposes the development and evaluation of a wearable technological device to acquire in situ the biomechanical variables of the operator, and by means of this data in conjunction with machine learning algorithms, it is possible to determine automatically and objectively the occupational risk indexes. It should be noted that the implementation of intelligent algorithms that estimate risk indicators based on biomechanical signals will be part of a future work.

This study is distributed as follows: Sect. 2 describes the hardware architecture of the wearable device and the location of the sensors on the operator; Sect. 3 deals with the implementation of the device, the software architecture of the embedded system, and the user interface; Sect. 4 documents the evaluation of the wearable device; finally, Sect. 5 summarizes the conclusions of the study.

2 Hardware Architecture of the Wearable Device

The hardware architecture of the wearable device is shown in Fig. 1. This architecture is divided into four modules: The signal acquisition module, which performs the function of capturing body and biomechanical signals through sensors. Processing module, this module consists of a microcontroller device that performs operations with the acquired data. Storage and Transmission Modules, which are formed by peripheral electronic devices to store and transmit the time series of data. Finally, this hardware architecture is connected to the Visualization Module, which is formed by a software interface whose main functions are to visualize and register the acquired data sets in files.

2.1 Signal Acquisition Module

The developed wearable device acquires 11 signals, the types of signals acquired along with the sensor used are mentioned below: myoelectric signals at the forearm, biceps, and back (Gravity EMG SEN0240 Sensor), back inclination angles (IMU MPU 6050 Sensor), heart rate (Pulse Sensor), skin electrical resistance/conductance signal (Grove GSR Sensor), and, body temperature (MLX90614 Sensor).

Myoelectric signal sensor. The OYMotion Gravity EMG SEN0240 sensor captures muscle activation signals employing metallic surface electrodes, known as myoelectric signals. The sensor has an integrated signal amplification and filtering circuit to increase the Signal to Noise ratio. The main features of this device are mentioned below: Detection Range: $\pm 1.5\text{mV}$, Output Voltage: 0–3.0 V, Operating Temperature: 0–50 °C, Electrode type: metal dry [8]. For the placement of electrodes on the user's body, specific locations and minimum distances between electrodes must be taken into account [9].

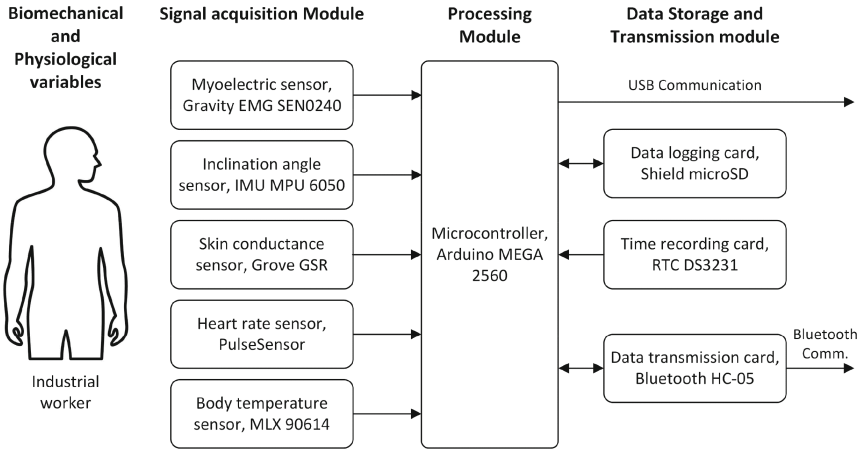


Fig. 1. Hardware architecture of the wearable device.

Inclination angle sensor. The IMU MPU 6050 sensor is an electronic device consisting of a 3-axis accelerometer, a 3-axis gyroscope, and a magnetometer. The sensor data is processed by a DMP (Digital Motion Processor) which applies complex data integration algorithms called 9-axis Motion Fusion. This sensor can be used to measure acceleration, velocity, orientation, displacement, and other motion-related parameters [10].

Skin conductance sensor. The user subjected to states of great physical exertion or emotional stress stimulates the sympathetic nervous system and consequently activates the sweat glands, which varies the skin conductance. The method for measuring the electrical conductance of the skin is called GSR (Galvanic Skin Response) and consists of characterizing the conductance using two surface Nickel electrodes placed at two specific points, for example, two fingers of the hand. The system applies a voltage between the electrodes and captures the resulting current. The sensor used in the present work is the Grove GSR v1.2. The electronics board contains a signal conditioning circuit that includes the OpAmp LM324 with fast slew rate, and finally a passive low-pass RC filter in order to obtain the signal envelope [11].

Heart rate sensor. Pulse Sensor is an electronic device that combines an APDS-9008 photoelectric sensor, a signal amplifier based on the OpAmp MCP6001 and a network of RC filters for noise reduction. This small device uses the optical pulse detection technique called Photoplethysmogram, therefore, it is suggested to place it on the fingertip or earlobe, however, it can be worn on the wrist for measurement purposes by means of a wearable device. The manufacturer mentions that the sensor should be placed on a part of the body that has capillary tissue. The clamping pressure between the sensor and the skin should be Sweat Spot (neither too hard nor too soft), this will produce a clean signal with readable amplitude. In addition, it should be taken into account that the output measurement has a stabilization time of 5–10 s. The main features of this device are as follows: Maximum Current Draw < 4 mA, Output Voltage Range 0.3–5.5 V, Wavelength LED output/sensor input 565 nm/525 nm [12].

Body temperature sensor. The MLX90614 device is an infrared sensor used to measure body temperature without contact. It consists of the MLX81101 infrared detector, the MLX90302 signal conditioner, a low noise amplifier, a 17-bit ADC, and finally the DSP that allows to obtain measurements with high accuracy and resolution. The sensor output is via 10-bit PWM (Pulse Width Modulated) and SMBus-I2C, the measuring range is from -20 to 120 °C, and the resolution is 0.14 °C [13].

2.2 Processing Module

The data processing module of the implemented device consists of the Arduino MEGA development board. The electronic board has the ATmega2560 microcontroller that operates at a frequency of 16 MHz and has 4 USART serial ports. The number of available ports (54 digital inputs/outputs and 16 analog inputs) permits the interconnection of the microcontroller with the number of sensors required in the present project [14].

2.3 Communication Module

The implemented prototype has the capacity to communicate wirelessly via Bluetooth. The electronic card that allows this connection is the HC-05 Module. Among the most relevant features are the following: Typical -80 dBm sensitivity, ISM 2.4 GHz frequency band, nominal range 10 m, GFSK modulation (Gaussian Frequency Shift Keying), and finally, the configuration of the operating parameters is compatible with AT commands [15].

2.4 Storage Module

Data logging card. This module has the main function of managing the internal logging of the data captured by the sensors. The electronic card that records the information inside the implemented prototype is the Arduino Shield microSD. The module supports two types of memory: SD or SDSC (Standard Capacity) and SDHC (High Capacity), and the communication protocol is SPI. This module has a library that allows to create, read and write files, it also has a physical socket in which the microSD memory containing the created files is inserted [16].

Time reading card. The implemented prototype requires the creation and registration of time series with the measured data. These time series need the date and time information in which the data are captured. The DS3231 real-time clock (RTC) card allows the action of generating the time labels (seconds, minutes, hours, day, month, and year) for the data. The communication protocol is I2C and it has an integrated battery for continuous time control [17].

2.5 Location of Sensors

The implemented device captures body signals which will provide information about the biomechanics of the user performing a work activity. This section documents the

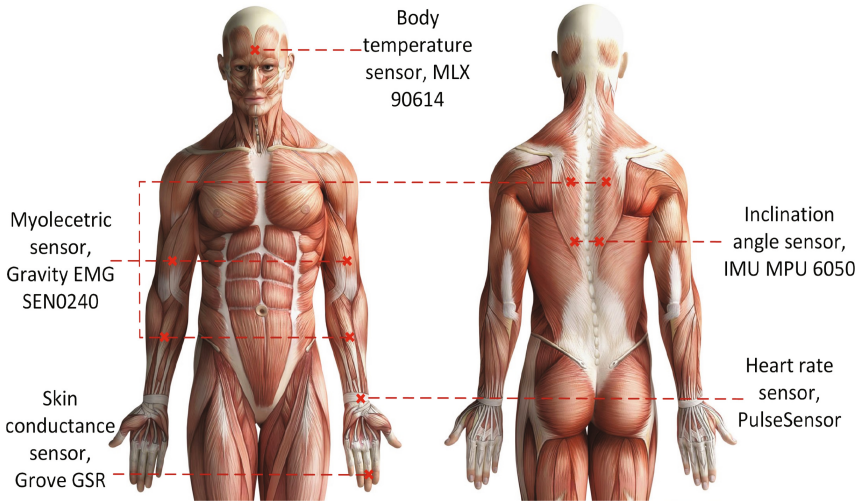


Fig. 2. Location of the sensors on the user of the wearable device.

location of each of the sensors. Figure 2 shows the location of the sensors on the user of the wearable device.

EMG sensor, forearm muscle. The electrode is placed in the middle third of the forearm, this gives us a myoelectric signal of greater amplitude according to the recommendation of the SENIAM Project. The SENIAM project (Surface ElectroMyoGraphy for the NonInvasive Assessment of Muscles) is a European concerted action in the Biomedical Health and Research Program (BIOMED II) of the European Union [18, 19].

EMG sensor, biceps brachii muscle. Based on the SENIAM recommendation, the electrode should be placed in the distal third of the biceps.

EMG sensor, middle trapezius muscle. The electrode should be placed in the portion closest to the spine of the middle trapezius muscle.

Inertial sensor, lower trapezius muscle. This sensor is used to examine the movement of the back, mainly rotation angles and movement frequencies. Accordingly, the sensor should be located in the left and right lower trapezius muscle.

GSR sensor, fingers. The measurement of skin conductance is obtained by placing two dry electrodes, it is recommended that the electrodes be placed on the middle and ring fingers.

Heart rate sensor, wrist. The sensor should be placed in the wrist area, as close to the radial artery as possible.

Body temperature sensor, forehead. The infrared temperature sensor can be located in some possible points of the user's body; however, the sensor's manual recommends its location on the forehead for greater reliability of the reading.

3 Wearable Device Implementation

3.1 Electronic Diagram

The electronic wiring diagram of the wearable device is presented in Fig. 3. This diagram takes into account the interconnection between the four modules of the system: data acquisition module, processing module, storage module and data transmission module. The diagram was documented in the Fritzing tool version 0.9.9.

3.2 Device Software Architecture

The algorithms executed in the device have the function of digitizing, conditioning, storing and transmitting the data captured by the sensors. These algorithms are structured in dedicated subroutines for each type of signal, in addition to the libraries provided by the sensor manufacturers. Figure 4 shows the block diagram of the device software architecture of the wearable device.

The “WProgram” and “EMGFilters” libraries are used to initialize the six sensors, read and filter the acquired data. The sampling frequency of these data is up to 500 samples per second and the notch filter cutoff frequency is 60 Hz to eliminate the noise coming from the electrical network that adds to the EMG signals captured by the electrodes. Changes were made to the original manufacturer’s library, mainly in the second and fourth order filters section, the nominal frequency band to be allowed is between 20 and 150 Hz, this is done to obtain the envelope of the myoelectric signal. Finally, the data transfer rate is defined at 9600 bps.

The “MPU6050” and “I2C” libraries are used to capture the data from the two inertial sensors. First the data generated by the accelerometers in the three axes (X, Y, Z) are read, then the inclination angles of the two sensors are determined.

For the reading of the GSR sensor no library is used, however to obtain the skin resistance (SR), Eq. (1) is used [20].

$$SR = \frac{(2^{10} + 2x)*R}{2^9 - x} \quad (1)$$

where x is the serial port reading, it should be noted that the reading range of the serial port is from 0 to 1023 due to the fact that the analog input has a 10-bit analog-to-digital converter. Finally, the resistance of the voltage divider circuit is $R = 10 \text{ k}\Omega$.

3.3 Software Architecture of the User Interface

The hardware device that captures the biomechanical signals sends the data to a user interface for real-time display and recording of information in CSV files. The user interface was developed in the Python programming language. The software architecture of the user interface is shown in Fig. 4.

The software architecture starts with the import of the “PySerial”, “Matplotlib” and “Tkinter” libraries, these libraries were used for serial communication and display of scalar and vector data. The received data frame contains 13 fields: date, time, body

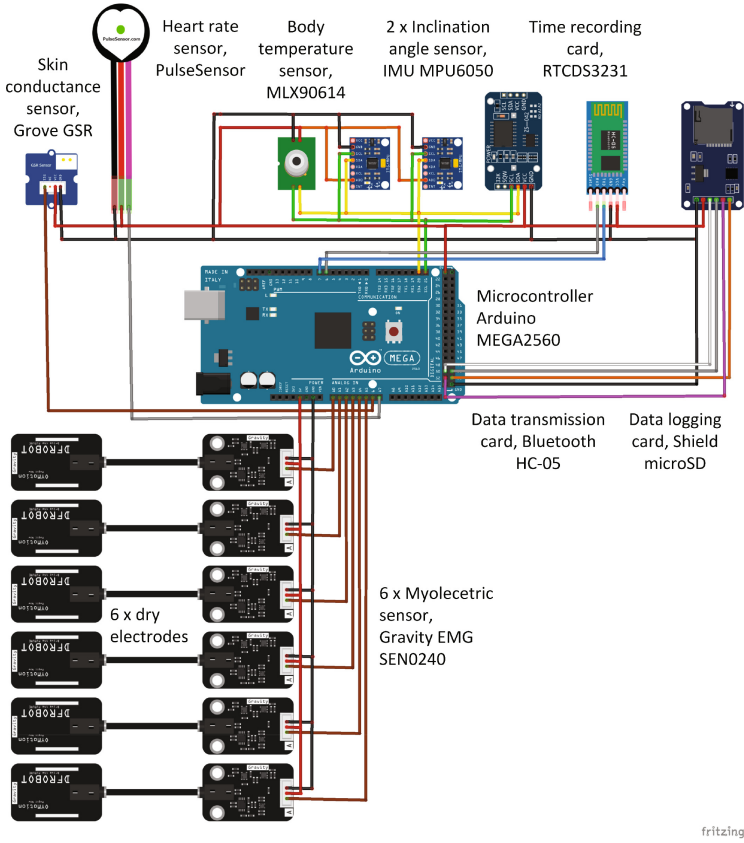


Fig. 3. Electronic wiring diagram of the implemented wearable device.

temperature, heart rate, left inclination angle, right inclination angle, skin resistance, left forearm EMG signal, right forearm EMG signal, left biceps EMG signal, right biceps EMG signal, left trapezius EMG signal and right trapezius EMG signal. Figure 5 shows the user interface developed in Python for displaying and registering information. The elements of the user interface are a selector of the signals to be displayed, a graph of the signal as a function of the acquired samples, instantaneous values of the monitored variables, date and time of the data reading, and finally data acquisition and display control buttons.

3.4 Wearable Device Integration

The main goal of the project is the in-situ measurement of biomechanical variables directed to the operating personnel in different industrial sectors. Therefore, this measuring device must have the characteristics of a Wearable. The Wearable implemented was based on the structure of a posture corrector vest available in the market aimed at occupational health and safety.

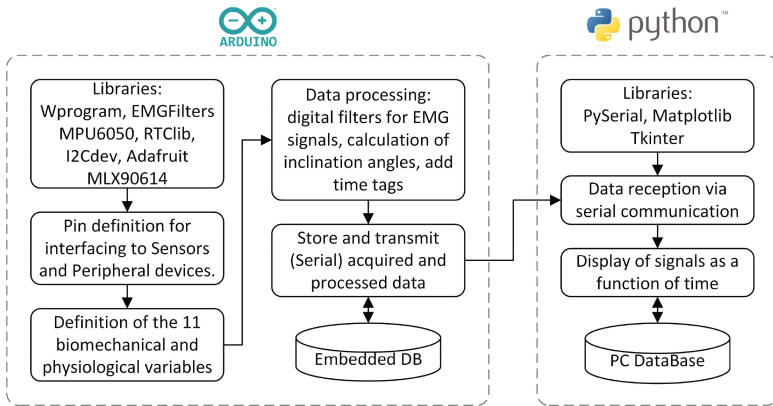


Fig. 4. Software architecture of the wearable device and user interface.

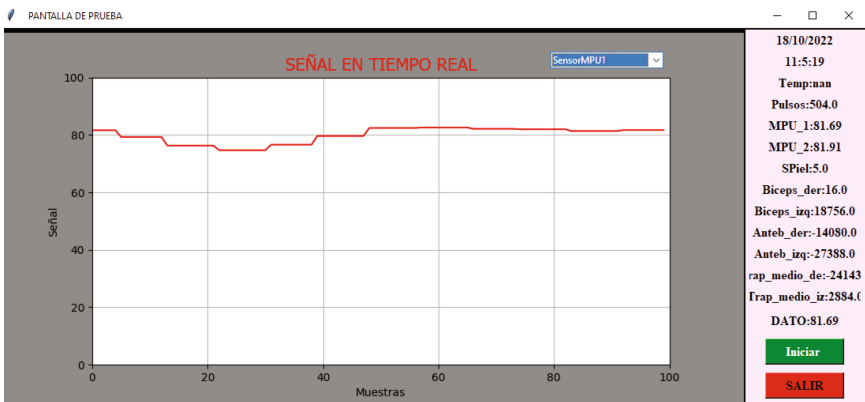


Fig. 5. User interface for real-time data visualization and data logging.

Figure 6 illustrates the exterior and interior view of the implemented Wearable device. Figure 6a of the external view of the Wearable consists of the following elements: (a) six Gravity EMG SEN0240 sensors to measure myoelectric signals in the arm, biceps, and back, (b) a Pulse Sensor to measure heart rate, (c) a MLX90614 sensor to measure body temperature, (d) a Grove GSR sensor to measure skin conductance, and, (e) posture corrector vest type garment. Figure 6b of the internal view of the Wearable shows the following elements: f) two IMU MPU 6050 sensors to measure the back inclination angles, g) the MicroSD module to store the information, (h) the Arduino Mega 2560 microcontroller, (i) the Bluetooth HC-05 module, (j) the RTC DS3231 module, and, (k) the power supply module for the electronic cards. It should be noted that the electronic cards are arranged in an orderly fashion inside the compartment located on the back outside of the vest, as shown in Fig. 7.

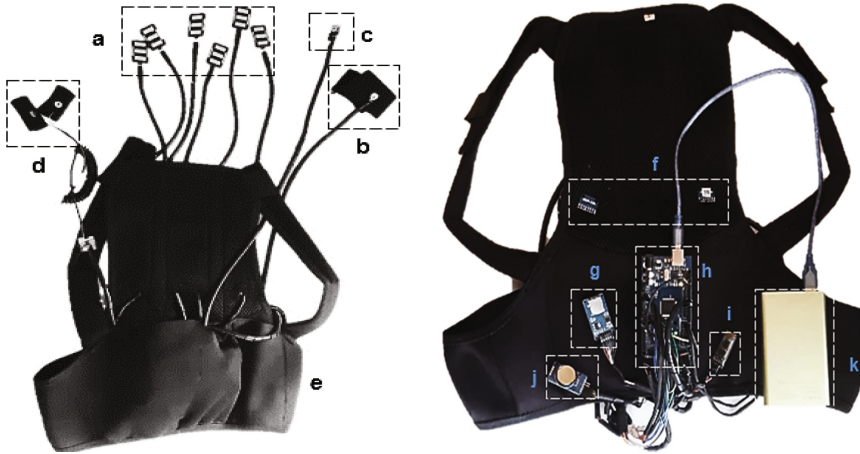


Fig. 6. Exterior (left) and interior (right) view of the Wearable device for measuring biomechanical variables.

4 Wearable Device Evaluation

4.1 Experimental Evaluation Methodology

The wearable device is able to acquire, store and transmit 11 signals from the sensors and also two data (time and date) that identify each of the acquired data frames. The evaluation tests of the implemented wearable device were performed with three users with different physiological parameters. The characteristics of the test users are summarized in Table 1.

Table 1. Features of the test users for the evaluation of the Wearable.

User	Height (m)	Weight (kg)	Age	Body mass index (BMI)
User 1	1.55	48	30	19.97, underweight
User 2	1.51	66	32	28.94, overweight
User 3	1.58	60	20	22.58, normal weight

Once the test users were defined, the wearable device was evaluated. The experimental setup consists of the wearable device implanted in each of the users, a computer to run the user interface, and the different test weights. The realization of the exercises or movements has a duration of less than 20 s. Figure 7 shows a picture of a test user using the wearable device.

For the definition of the test exercises to be executed by the users, the OWAS (Ovako Work Posture Analysis System) method was used as a reference, which is one of the most complete postural evaluation systems for analyzing the risks derived from forced postures in the execution of tasks [6].



Fig. 7. Test user using the implemented wearable device.

Three exercises were defined for the test scenario: In exercise 1 the test user lifts a 9 kg load from the floor to a height of 70–80 cm. In exercise 2 the user's back is bent at an angle of 20° with a load of 9 kg. In exercise 3 the user must lift a 9 kg load to a height of approximately 120–130 cm. A photograph of the three test exercises performed by the user is presented in Fig. 8.

Once the three test exercises have been defined, the posture type is verified using the OWAS method. Table 2 shows the posture codes based on the OWAS method for the test exercises.



Fig. 8. User executing the three test exercises for the evaluation of the wearable device.

4.2 Data Obtained by the Device

The wearable device acquired and registered the eleven variables to characterize the postural activity executed by the user. The execution of the three test exercises by the three users generated a set of 99 signals. The time interval of the signals ranged from 9 to 11 s, depending on the execution of the exercise by the test users. Figure 9 shows the

Table 2. Posture codes based on the OWAS method for the three test exercises.

Item	Back code	Arms code	Legs code	Load
<i>Exercise 1</i>				
Code	1	1	2	1
Posture	Straight back	Below shoulders	Standing up	< 10 kg
Risk	1	Posture is normal and natural so no action is required		
<i>Exercise 2</i>				
Code	2	1	2	1
Posture	Inclined back > 20°	Below shoulders	Standing up	< 10 kg
Risk	2	Posture can cause damage and requires action at a future period		
<i>Exercise 3</i>				
Code	1	2	2	1
Posture	Straight back	Above the shoulders	Standing up	< 10kg
Risk	1	Posture is normal and natural so no action is required		

11 signals acquired by the Wearable device; the signals correspond to user 2 executing exercise 3.

From this set of data, a general analysis of the coherence of the variables is made, for example, there are specific moments where a significant increase in the electromyographic activity of various muscles is observed, suggesting periods of intense physical exertion. These periods are also accompanied by increases in heart rate, which is an expected physiological response to physical exertion. Back tilt also varies, most likely due to changes in posture during these periods of exertion.

The following is a brief analysis of the acquired set of signals:

The signals captured by the EMG sensors exhibit a sequence of signal peaks coherent with the times and intensities of muscle contraction accompanied by the stretching of the muscle fibers that produce the movement; consequently, an eccentric contraction is obtained. In addition, low voltage levels are observed corresponding to the time intervals in which there is muscle relaxation. In exercise 3, signals with greater amplitude and a greater number of signal peaks were identified, this is due to concentric muscle contraction.

The signals captured by the inertial sensors correspond to the inclination angles of the back (left and right sides). Even though the back constantly changes its angle of inclination due to the constant movements of the user, the noise contained in the signals is minimal. In the signals of exercises 1 and 3 (perpendicular back), the average angle per user ranges between 79° and 85°, in the signal of exercise 2 (inclined back) the average angle per user ranges between 60° and 70°, which is congruent with the actual posture of the user.

With respect to the skin conductance signals, it is observed coherence between the highest amplitudes of the signal with the time intervals where a major physical effort

is generated. The lowest amplitude levels range between 5 and 6 μV and the highest amplitude levels range between 8 and 10 μV . The intention is to use this signal to establish a relationship with the fatigue and stress of the user executing a task, this relationship will be strengthened with the heart rate and body temperature signals.

The range of temperature levels acquired by the sensors is between 35 and 38 $^{\circ}\text{C}$, which is consistent with the World Health Organization's (WHO) statement that normal body temperature measured at the forehead is between 35.4 and 37.4 $^{\circ}\text{C}$. The temperature signals exhibit curves with slight increases and decreases in amplitude with respect to the average value, these changes between the beginning and the end of the exercise are approximately 1 $^{\circ}\text{C}$.

5 Conclusions

A wearable device based on IoMT (Internet of Medical Things) elements applied to the in-situ measurement of biomechanical variables was designed and implemented. This is a portable and low-cost device focused on capturing 11 variables for the analysis of occupational risks in industrial environments. It is necessary to emphasize that the device meets the functional requirements, however, it is still considered a test device in a controlled environment and not as a professional medical equipment.

The hardware architecture of the implemented wearable device consists of four modules: signal acquisition module, processing module, storage module, and data transmission module. The sensors to measure the 11 variables are the following: 6 Gravity EMG SEN0240 sensors to measure myoelectric signals in the forearm, biceps, and back, 2 IMU MPU 6050 sensors to measure back inclination angles, a Pulse Sensor to measure heart rate, a Grove GSR sensor to measure skin electrical resistance/conductance, and, a MLX90614 sensor to measure body temperature.

To evaluate the performance of the implemented device, the 11 monitored variables were acquired, recorded, and visualized during the execution of three test exercises by three users of different physiological parameters, resulting in a set of 99 signals with a duration of 9–11 s, depending on the execution of the exercise by the user. When analyzing this initial set of acquired signals, it is observed that there is no loss of data between capture, transmission, recording, and visualization, in addition, congruence is observed between the occurrence of peaks of the signals with the features of the executed exercises.

One of the main limitations of the proposed system is the effect of the constant movement of the operator that alters the adhesion between the electrodes and the skin. These frictions will generate changes in the signal and therefore errors in the measurements. For this reason, from the beginning of the project, signal acquisition prioritizes the signal envelope instead of the complete signal.

The next phase of the project consists of developing and training a machine learning system, based on supervised learning [21] or reinforcement learning [22], to combine the data obtained and determine risk indicators for a defined number of occupational injuries. In addition, the acquired data set must be analyzed and labeled based on expert and evidence-based criteria.

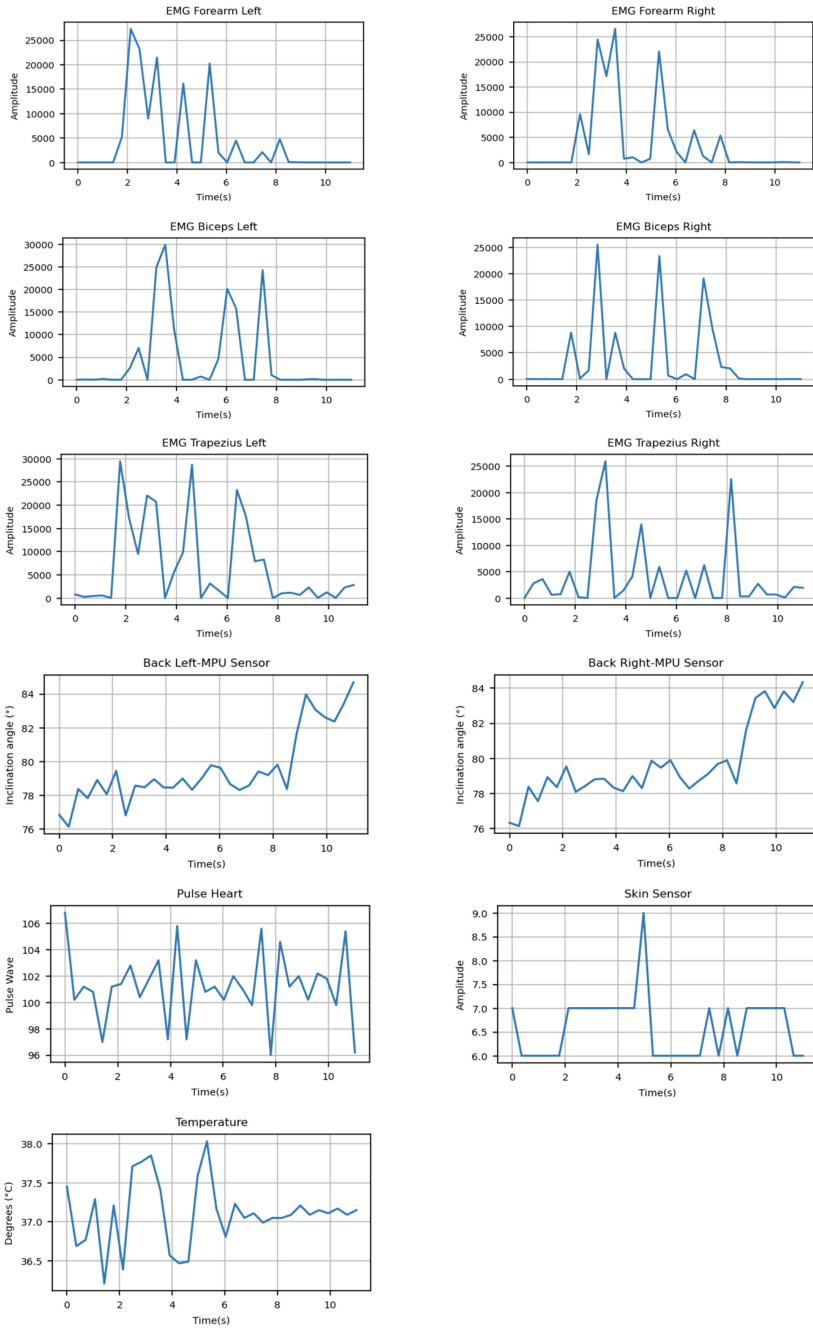


Fig. 9. Signals acquired by the wearable device: 11 variables of user 2 performing exercise 3.



References

1. Salvador, L.R., Van Thinh, D.: Occupational safety and health: an overview. In: 2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 355–360. IEEE (2016)
2. Jiang, Z., Bakker, O.J., Bartolo, P.J.: Critical review of Industry 4.0 technologies' applications on occupational safety and health. In: 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), vol. 1, pp. 1267–1272. IEEE (2022)
3. Jafarian, M., Hoseini, N., Safahieh, A., Namnabat, M.S., Arjmand, N.: Comparison of different risk assessment tools to manage musculoskeletal disorders among workers in an automobile manufacturing company. In: 2023 30th National and 8th International Iranian Conference on Biomedical Engineering (ICBME), pp. 253–259. IEEE (2023)
4. Iqbal, M., Angriani, L., Hasanuddin, I., Erwan, F., Soewardi, H., Hassan, A.: Working posture analysis of wall building activities in construction works using the OWAS method. *IOP Conf. Ser. Mater. Sci. Eng.* **1082**(1), 012008 (2021)
5. Dzeng, R.-J., Hsueh, H.-H., Ho, C.: Automated posture assessment for construction workers. In: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1027–1031. IEEE (2017)
6. Jankovic, P., Protic, M., Jovanovic, L., Bacanin, N., Zivkovic, M., Kaljevic, J.: YOLOv8 Utilization in Occupational Health and Safety. In: 2024 Zooming Innovation in Consumer Technologies Conference (ZINC), pp. 182–187. IEEE (2024)
7. Di Pasquale, V., De Simone, V., Radano, M., Miranda, S.: Wearable devices for health and safety in production systems: a literature review. *IFAC PapersOnLine* **55**(10), 341–346 (2022)
8. DFROBOT: Analog EMG Sensor by OYMotion, 12 July 2024. Available: https://wiki.dfrobot.com/Analog_EMG_Sensor_by_OYMotion_SKU_SEN0240
9. D. G. d. I. Biomedica: Frecuencia Cardíaca, 12 July 2024. Available: <https://www.dalcame.com/fc.html>
10. Hassan, A., et al.: Statistical scheme for fault detection using Arduino and MPU 6050. In: 2019 Prognostics and System Health Management Conference (PHM-Qingdao), pp. 1–7. IEEE (2019)
11. S. Studio: Grove—GSR, 1st edn
12. Gitman, Y., Murphy, J.: Heartbeat Sensor Projects with PulseSensor: Prototyping Devices with Biofeedback, 1st edn. Apress, Berkeley, CA (2023). <https://doi.org/10.1007/978-1-4842-9325-6>
13. M. I. w. Heart: Datasheet for MLX90614, 12 July 2024. Available: <https://www.melexis.com/en/documents/documentation/datasheets/datasheet-mlx90614>
14. Arduino: Arduino® MEGA 2560 Rev3, 12 July 2024. Available: <https://docs.arduino.cc/resources/datasheets/A000067-datasheet.pdf>
15. I. S. M. I. easter: HC-05—Bluetooth to Serial Port Module, 12 July 2024. Available: https://components101.com/sites/default/files/component_datasheet/HC-05%20Datasheet.pdf
16. N. Mechatronics: Tutorial Arduino y memoria SD y micro SD, 12 July 2024. Available: https://naylampmechatronics.com/blog/38_tutorial-arduino-y-memoria-sd-y-micro-sd.html
17. A. Devices: DS3231—Extremely Accurate I2C—Integrated RTC/TCXO/Crystal, 12 July, 2024. Available: <https://www.analog.com/media/en/technical-documentation/data-sheets/ds3231.pdf>
18. Gonzalez-Ibarra, J.C., Soubervielle-Montalvo, C., Vital-Ochoa, O., Perez-Gonzalez, H.G.: EMG pattern recognition system based on neural networks. In: 2012 11th Mexican International Conference on Artificial Intelligence, San Luis Potos, Mexico, pp. 71–74 (2012)
19. Hazam Majid, M.S., Khairunizam, W., Sahyudi, B.N., Zunaidi, I., Shahrman, A., Zuradzman, M.: Determining acceptable range of surface electromyogram electrode placement variation

- for deltoid muscle using euclidean distance function. In: 2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA), Kuching, Malaysia, pp. 1–5 (2018)
20. Iadarola, G., Bruschi, V., Cecchi, S., Dourou, N.A., Spinsante, S.: Low-cost monitoring for stimulus detection in skin conductance. *Acta IMEKO* **12**(3) (2023)
 21. Calderon, C., Ramírez, C., Barros, V., Quezada-Sarmiento, P.A., Barba-Guamán, L.: EMG signal patterns recognition based on feedforward artificial neural network applied to robotic prosthesis myoelectric control. In: 2016 Future Technologies Conference (FTC), San Francisco, CA, USA, pp. 868–875 (2016)
 22. Calderón-Cordova, C., Sarango, R., Castillo, D., Lakshminarayanan, V.: A deep reinforcement learning framework for control of robotic manipulators in simulated environments. *IEEE Access* **12**, 103133–103161 (2024)



Spectral Analysis of Powertrain Vibration in a Hybrid Vehicle Under Controlled Operating Conditions

Raquel de los A. Salas Ibarra¹(✉) , Alexander E. Torres Romero^{2,3} ,
and David H. Cárdenas Villacrés³ 

¹ Telecommunication Systems Research Group (GISTEL), Universidad Politécnica Salesiana, Campus Centenario, Guayaquil, Ecuador

{rsalasi, atorresr, dcardenasv}@ups.edu.ec

² Electronic and Automation Engineering, Sustainability, Management and Regulation of Telecommunications and Energy Research Group (SMART-TECH), Universidad Politécnica Salesiana, Campus Centenario, Guayaquil, Ecuador

³ Telecommunication Engineering, Telecommunication Systems Research Group (GISTEL), Universidad Politécnica Salesiana, Campus Centenario, Guayaquil, Ecuador

Abstract. This paper deals with the characterization of the vibration frequency spectrum of the electric motor in the power train of a hybrid vehicle. This analysis is carried out using a vibration analyzer as a measurement instrument and Minitab software for processing the numerical data obtained. The numerical data collected in each experiment allow an accurate evaluation of the significant differences in power consumption, both in the charging and discharging phases. In order to meet the proposed objectives, we proceed with the characterization of the vibration spectrum of the electric motor of the hybrid vehicle powertrain under controlled conditions, which include both the charging and discharging phases, as well as the operation of the air conditioning system at its maximum capacity. The acquisition of this data is carried out using Dewesoft software. Subsequently, analysis of variance is used to process and filter the collected data. The null hypothesis was accepted, since the results exceeded the significance level set at 0.005, with a confidence interval of 99.5% for normal distribution and 99% for ANOVA analysis.

Keywords: Spectrum · Frequency · Johnson Transform · Power Train · ANOVA · FFT

1 Introduction

Improved energy efficiency and reduced exhaust emissions have been the main drivers for the rapid development of hybrid electric vehicle (HEV) technologies. The first wave of commercial HEV offerings combined the flexibility of internal combustion engines (ICE) with all-electric traction systems to provide optimal

vehicle operation. Advanced control methodologies employing rules based on the optimization of fuel efficiency and no emissions have been developed in this area. Despite this progress, and the many components of an HEV vehicle powertrain, the characterization of vibration sources in a hybrid vehicle and the propagation of vibration in structural response and interior cavities has not yet been fully addressed. It is expected that this issue will become even more important for the coming generation of vehicle structures and in particular for electric vehicles where the acoustic comfort of the vehicle interior will play a greater role in occupant perceived quality [1, 17].

The powertrain systems generating forced and/or self-excited vibrations are the engine, the electric motor(s), and the continuously variable transmission (CVT). Even if the overall level of mechanical noise inside the cabin from these systems can be well controlled by passive means by vibration isolation and damping solutions at their interfaces with the vehicle, specific spectral bands can be identified that could potentially excite natural resonant modes of the vehicle or of the floor. In this respect, battery power management and operation of the electric motors in mode conditions can represent a critical hazard considering that the boom phenomenon could also happen [11, 14].

A hybrid vehicle consists of two or more energy supply systems that can supply power to the powertrain independently or together [12, 15]. These vehicles combine an internal combustion engine with an electric motor, resulting in lower fuel consumption and reduced pollution [3].

There are three powertrain configurations for hybrid vehicles: series hybrid, parallel hybrid and series-parallel hybrid [5, 8]. The series-parallel hybrid combines the advantages of the two configurations, allowing the vehicle to be powered by only one of the engines or by both, depending on the driving modes [16, 20]. The vehicle studied has a series-parallel configuration, which allows versatile operation with characteristics of both configurations [2].

The operation of hybrid vehicles depends significantly on electric motors, which must have high power density, efficiency and reliability. Over the years, electric motors have evolved and have been used in both industry and vehicles [10]. The most common types in the automotive industry are brushless direct current (DC) and alternating current (AC) motors, which include permanent magnets, induction and switched reluctance. The classification of motors is made according to their nature into DC and AC, according to the NEMA standard [15].

Predictive maintenance uses specialized techniques to predict failures in equipment and facilities, some techniques include vibration analysis, ultrasound analysis, lubricating oil analysis and thermography [18]. Vibration analysis is based on the measurement of the repetitive motion of a body around its equilibrium position. Vibration measurement is crucial for evaluating machine performance and detecting possible failures [19].

Likewise, this technique is also based on frequency analysis which is measured through its amplitude, thus when the vibration spectrum presents variations throughout the measured operating period, it indicates that the machine presents an alteration in its operation and needs to be inspected [13]. Frequency analysis

is a technique used to evaluate the vibration spectrum and detect alterations in the machine operation, which allows identifying the need for inspection and maintenance [19].

According to Roland [9], he mentions that electric mobility has increased worldwide, with 1.82 million BEVs and 0.83 million PHEVs sold by 2021, 168% more than in 2020. Similarly, the AEADE reports that Ecuador has increased the acquisition of these units due to the environmental impact that this generates in a positive way. Between 2017 and 2020 an average of 115 EVs were sold. More the EVs had more reception in 2021 approximately 4,261 vehicles were sold, four times more than the previous year. In the months of January and February, approximately 883 units were sold, twice more than what was sold in 2021 according to the AEADE. This led to a high demand for HEVs in the national vehicle fleet, thus generating the need for maintenance of this equipment in order to achieve reliability and maintainability, according to Mesara et al. [6].

Maria Penkova [5], mentioned that maintenance management for a long time has been limited, but due to industrial growth and acquisition of technological equipment there is a need to implement it. Predictive maintenance was the next step to technology, founded to detect failures before they happen, the most significant vibration analysis and concludes that this technique allows detecting problems during installation, operation and final stage of the equipment, which contributes to a comprehensive control of maintenance costs and productivity improvement, mentioned Gallará and Pontelli [4].

In addition, Swapnil and Prasad [7], emphasize a comprehensive overview of condition monitoring and fault diagnosis techniques in hybrid vehicles. They also mention that vibration spectral analysis is a key tool for the use of hybrid vehicle powertrain monitoring. It is important the application of predictive maintenance techniques in electric and hybrid vehicles.

2 Materials and Methods

For the methodology proposed for the realization of the vibration data collection provided by the electric motor of the hybrid electric vehicle power train. The following considerations were taken into account with respect to testing, such as: requirements (test elements), preconditions for data collection, test protocol and the ISO 10816-3 standard, on which this proposed methodology is based.

2.1 Test Process

The experimental procedure is developed following a series of steps defined in the flow diagram as shown in Fig. 1. The accelerometer is placed as identified in the diagram in Fig. 2, which represents the distribution of the vehicle's electric motor. To start the experimental recording, the vehicle is turned on, verifying that the air conditioning system is at maximum, with a sampling time of 10 s, repeating the sampling three times, concluding with the processing of the data using the Dawesoft software.

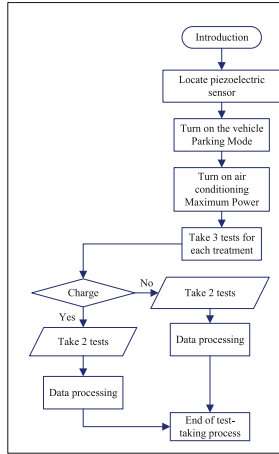


Fig. 1. Test protocol flowchart.

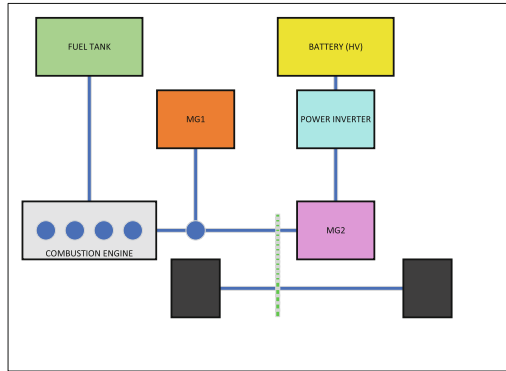


Fig. 2. Location of the piezoelectric sensor.

2.2 Dewesoft Equipment Resolution Setting

The vibration analyzer is connected to the computer through a network cable. By means of the Dewesoft software, which has the corresponding license, the vibration data of the electric motor is recorded, both in load and unload mode, these data are censored with the help of the piezoelectric sensor. After the piezoelectric sensor is placed on the electric motor, the Dewesoft software is configured for data acquisition. Consequently, the analog input is configured, with a bandwidth of 100 Hz, corresponding to the sampling frequency.

IEPE is selected which refers to the piezoelectric accelerometer sensor, with a setting of 1000 mV, after the analog input configuration, we go to setting. In the default sensor configuration, the unit of measurement is in m/s^2 . However, to achieve an accurate conversion, a sensitivity factor of 100 will be applied when

the “g” (gravity) option is selected. Subsequently, the unit of m/s^2 is selected again, and the software provides us with the calculated sensitivity factor, which is equivalent to 10.1972 mV/m/s^2 .

After configuring the mathematical part of the software, we proceed to collect data in order to obtain the vibration spectra in frequency. In the same section of “Add mathematics”, the FFT (Fast Fourier Transform) is selected, this is located in the section of “Analysis in the frequency domain”, thus completing the total configuration of the equipment.

2.3 Minitab Software Configuration

To perform statistical data processing, it is essential to configure it correctly according to the needs of the project. The process involves:

- Start the software and select “Johnson Transform” under statistics and quality tools.
- Configure the Johnson transform by adjusting parameters, such as the column to transform.
- This transformation seeks to make the data follow a normal distribution.
- Apply the transformation to samples in both loading and unloading process.
- Perform a normality test to check if the transformed data are normal.
- Perform an ANOVA to identify significant differences in the amplitude data.
- Employ a 95% confidence level and the Tukey method to control for errors in mean comparisons.

This process allows amplitude data to be analyzed and significant differences in spectral data to be reliably detected.

3 Results and Discussion

The behavior of variables obtained through data collection and their application to statistics using Minitab software is analyzed. The Johnson transform was used to normalize the data of the frequency and amplitude variables, collected by the Dewesoft software, which uses a piezoelectric sensor to measure the vibrations of the power train during the loading and unloading processes under energy consumption conditions. Prior to collecting the data, the resolution of the Dewesoft vibration analyzer equipment was configured to select the vibration frequency spectrum for both the loading and unloading processes. This involved setting the sampling frequency and defining a time range for data acquisition per second. As a result, a resolution of 1024 lines over a period of 10 s was obtained, with a sampling frequency of 1 kHz. The graphs obtained in the treatments and trials showed differences, but not significant differences in the means. The null hypothesis was accepted because it exceeded the established significance level of 0.005. The results were ratified by means of the Tukey test, confirming that there are no significant differences between the variables analyzed in the loading and unloading processes, as well as in the number of trials carried out with a 99% confidence level.

3.1 Energy Consumption of the Loading Process - Test 1; Interpretation of the Statistical Method Johnson Transform and Normality Test

Figure 3 shows the Johnson transform of the first test in the loading process with the engine running and the air conditioning system at maximum. A confidence interval of 0.005 was used to convert the non-normal data to normality and analyze their distribution and correlation. The null hypothesis is accepted with a “p” value of 0.239, indicating that there are no significant differences in the transformed data.

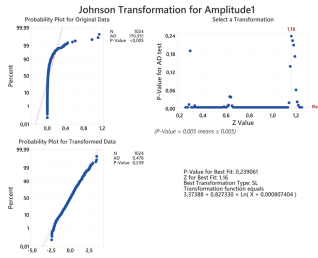


Fig. 3. Johnson transformation sample 1 loading process (trial 1).

In relation to Fig. 4, the normality test shows that the transformed numerical data have a normal distribution, i.e. the mean, mode and median coincide, thus making it a symmetrical distribution that is reliable for generalizing and inferring the sample as a general population.

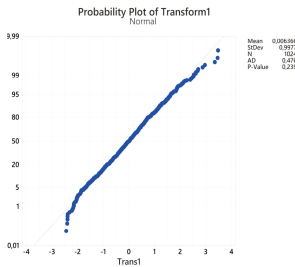


Fig. 4. Normality test sample 1 process load (trial 1).

3.2 Interpretation of the Analysis of Variance Method (ANOVA) Single Factor and the Composite Spectrum

Statistical graphs are presented below to verify significant differences. The number of factors evaluated using the Analysis of Variance (ANOVA) statistical method is presented generally in Table 1 and specifically in Table 2. Only one factor was evaluated with data transformed to normality.

Table 1. Number of factors to be analyzed (test 1).

Factor	Levels	Values
Factor 4	4	Transformation 1; Transformation 2; Transformation 3; Transformation 4

Tukey's method was applied in ANOVA with a 95% confidence interval to analyze significant differences in the grouped data, as can be seen in Fig. 5. This method was used because of the need for multiple comparisons in the trials.

The grouping of the means and it is concluded that, although there is a small difference in transformants 3, 4, 2 and 1, they are not significant and are considered as group A, as seen in Table 2. Therefore, it is determined that the means in trial 1 do not differ significantly.

Table 2. Grouping information using the Tukey method and 95% confidence

Factor	N	Mean	Grouping
Transformation 3	1024	0.0261	A
Transformation 4	1024	0.0117	A
Transformation 2	1024	0.0070	A
Transformation 1	1024	0.0064	A

The composite spectrum generated by Minitab software using the amplitude data of the test samples is depicted in Fig. 6. The amplitudes denoting changes in the Fourier transform in the "Block Count" section were selected. This is due to the variation in the numerical data in relation to machine vibration over time, also considering the resolution previously configured in the vibration analyzer equipment to evaluate the Fast Fourier Transform (FFT) in each test. Due to the lack of significant differences between the samples, it is concluded that samples 1, 2, 3 and 4 of the load test are relatively equivalent.

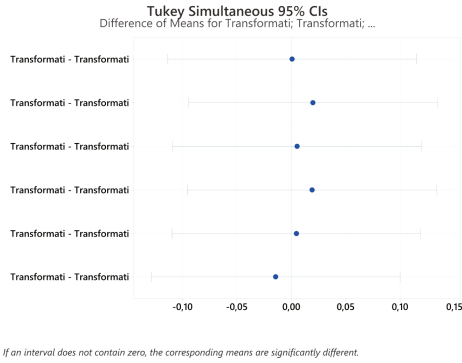


Fig. 5. Difference in the means in relation to the samples (trial 1).

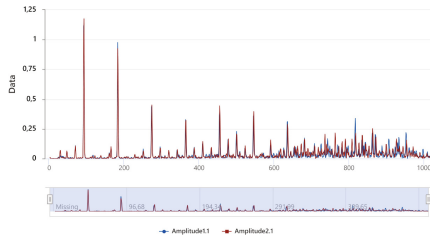


Fig. 6. Composite spectra of samples under load - time series plot (test 1).

3.3 Complete Treatment 1 - Charging Process with the Air Conditioning System at Maximum Air Conditioning

The superimposed spectra of tests 1 and 2 in the loading process are shown in Fig. 7. As explained above, since there are no significant differences in the means, a characteristic spectrum of each test is chosen in order to characterize the total treatment in the loading process.

In Table 3 shows the number of factors involved to be statistically analyzed; only the data transformed to normality will be analyzed for subsequent verification in ANOVA.

Table 3. Number of factors to be analyzed in loading process (test 1 and test 2).

Factor	Levels	Values
Factor 2	2	Transformation 1; Transformation 2

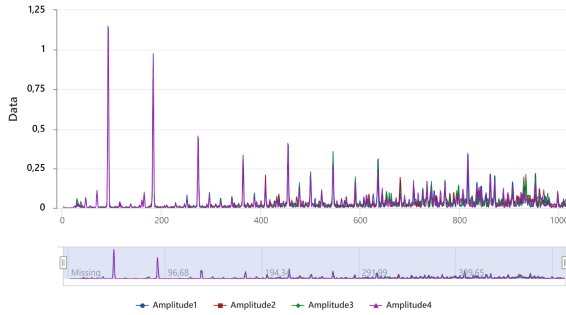


Fig. 7. Composite spectra of tests in loading process - time series graph (test 1 and test 2).

Tukey’s pairwise comparisons, which corresponds to a statistical method of ANOVA analysis, can be seen in Fig. 8. The correlation between trials 1 and 2, in which they correspond to the same group A, is detailed in Table 4. There are no significant differences since there are no differences in the means.

Table 4. Grouping information using the Tukey method and 95% confidence

Factor	N	Mean	Grouping
Transformation 1.1	1024	0.0064	A
Transformation 2.1	1024	-0.0069	A

3.4 Complete Treatment 2 - Unloading Process with the Air Conditioning System at Maximum Air Conditioning

The overlapping spectra of tests 1 and 2 in the unloading process are identified in Fig. 9. As in the loading process, no significant differences in the means were found. In this treatment, a characteristic spectrum of each test is selected to characterize the discharge process as a whole. Although the graph shows different peaks from each other, a statistical analysis will be performed to filter and verify if there are significant differences in the data.

The factors used in the statistical analysis of analysis of variance (ANOVA) are shown in Table 5. Values transformed to normality were selected to increase reliability in the analysis of the population sample.

The respective grouping carried out by the ANOVA statistical method is shown in Fig. 10, where the respective grouping of each transformed amplitude is detailed. In this case, they belong to the same group “A” as detailed in Table 6, which indicates that there are no significant differences in the means of the population samples, so it is possible to characterize the spectrum of both trial 1 and trial 2.

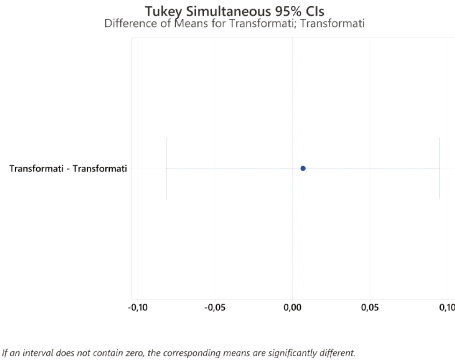


Fig. 8. Difference of the means in relation to the loading process samples (test 1 and test 2).

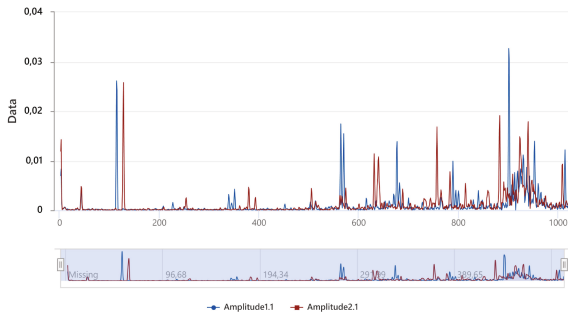


Fig. 9. Composite spectra of samples in discharge process - time series graph (test 1 and test 2).

Table 5. Number of factors to be analyzed in loading process (test 1 and test 2).

Factor	Levels	Values
Factor	2	Transformation 2.1; Transformation 3.1

Table 6. Tukey’s pairwise comparisons in download process (trial 1 and trial 2).

Factor	N	Mean	Grouping
Transformation 2.1	1024	-0.0111	A
Transformation 1.1	1024	-0.0179	A

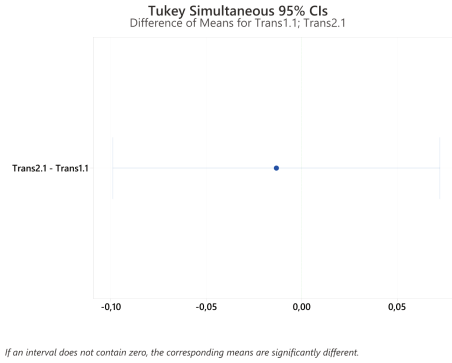


Fig. 10. Difference of means in relation to samples (trial 1 and trial 2).

4 Conclusions






- As part of this study, a comprehensive methodology for vibration assessment of the electric motor in the powertrain of a hybrid vehicle was designed based on the criteria established by ISO 10816-3. This meticulous approach included the implementation of rigorous procedures for data acquisition, strategic placement of sensors, precise duration of measurements, and consideration of the resolution of the measurement equipment. As a result of these practices, accurate numerical data on the vibrations generated by the electric motor of the powertrain were obtained in an informed manner, contributing to a detailed understanding of its vibrational behavior.
- According to the statistical methods previously mentioned, it was possible to obtain the characteristic frequency spectrum of each test in the energy consumption condition, both in the loading and unloading processes. With two trials for each treatment, no significant differences were obtained in the differences of the means, which leads to accept the null hypothesis since it is greater than the established significance level 0.005 with a confidence interval of 99.5% for the normal distribution and 99% for ANOVA analysis.
- In the analysis of the spectra generated by the electric motor vibrations, statistical methods were used to characterize each set. Techniques were applied to calculate the normal distribution of these spectra, followed by an analysis of variance (ANOVA) to evaluate the presence of significant differences between the means of different treatments. The results obtained indicate that, both in the loading and unloading processes, the spectrum of treatment 1 is substantially comparable to that of treatment 2. This finding is statistically supported by demonstrating that the spectra obtained are comparatively similar, supporting the hypothesis of equivalence between both treatments, since no significant differences were found to discredit it.

References

1. Bai, S., Liu, C.: Overview of energy harvesting and emission reduction technologies in hybrid electric vehicles. *Renew. Sustain. Energy Rev.* **147**, 111188 (2021)
2. Chris, M.: *Hybrid Electric Vehicles*
3. Ehsani, M., Gao, Y., Miller, J.M.: Hybrid electric vehicles: architecture and motor drives. *Proc. IEEE* **95**(4), 719–728 (2007)
4. Emadi, A.: 2007 IEEE Proceedings Modeling HEV **95**(4), 729–745 (2007)
5. Gao, D.W., Mi, C., Emadi, A.: Modeling and simulation of electric and hybrid vehicles. *Proc. IEEE* **95**(4), 729–745 (2007)
6. Grajales, D.H.M.: *Sci. Tech. Año XII* (2006)
7. Gundewar, S., Kane, P.: Condition monitoring and fault diagnosis of induction motor. *J. Vib. Eng. Technol.* **9**, 10 (2020)
8. Khan, A.: A comprehensive overview on the architecture of hybrid electric vehicles (HEV). In: *Proceedings of the IEEE*, pp 1–7 (2016)
9. King, N.: EVs Forecast to Account for Two Thirds of Global Light-Vehicle Sales in 2035
10. Liu, W.: *Hybrid Electric Vehicle System Modeling* (2017)
11. Masri, J., Amer, M., Salman, S., Ismail, M., Elsisi, M.: A survey of modern vehicle noise, vibration, and harshness: a state-of-the-art. *Ain Shams Eng. J.* 102957 (2024)
12. Osornio-Correa, C., Villarreal-Calva, R.C., Estavillo-Galsworthy, J., Molina-Cristóbal, A., Santillán-Gutiérrez, S.D.: Optimization of power train and control strategy of a hybrid electric vehicle for maximum energy economy. *Ing. Investig. Tecnol.* **14**(1), 65–80 (2013)
13. Pinos, R.J.A.: *Escuela Académico Profesional de Ingeniería Mecánica Desarrollo de plan de mantenimiento en motores eléctricos* (2019)
14. Qin, Y., et al.: Noise and vibration suppression in hybrid electric vehicles: state of the art and challenges. *Renew. Sustain. Energy Rev.* **124**, 109782 (2020)
15. Sanz, S.: *Vehículos híbridos y eléctricos. Motores*, p. 18 (2017)
16. Tran, D.D., Vafaiepour, M., El Baghdadi, M., Barrero, R., Van Mierlo, J., Hegazy, O.: Thorough state-of-the-art analysis of electric and hybrid vehicle powertrains: topologies and integrated energy management strategies. *Renew. Sustain. Energy Rev.* **119**, 109596 (2020)
17. Verma, S., et al.: A comprehensive review on energy storage in hybrid electric vehicle. *J. Traffic Transp. Eng. (Engl. Ed.)* **8**(5), 621–637 (2021)
18. Vicente, M.: *Mantenimiento Técnicas y aplicaciones*
19. White, G.: *Introducción al Análisis de Vibraciones*. In: *International Computer Music Conference*, p. 148 (2010)
20. Yilmaz, M., Krein, P.T.: Review of integrated charging methods for plug-in electric and hybrid vehicles. In: *2012 IEEE International Conference on Vehicular Electronics and Safety, ICVES 2012*, pp. 346–351 (2012)



Evaluation of an Intelligent Educational Toy-Game Prototype for Toddlers' Motor Stimulation and Learning

Nayeth Idalid Solorzano Alcivar^(✉) , Da Hee Park Kim ,
Jimmy Ernesto Canizares Pozo , Michael Xavier Arce Sierra ,
and Andrea Paola Rubio Zurita 

Escuela Superior Politécnica del Litoral, Campus Gustavo Galindo Km 30.5 Vía Perimetral,
Guayaquil, Ecuador
{nsolorza, dhpark, jecaniza, micxarce, anparubi}@espol.edu.ec
<http://www.espol.edu.ec>

Abstract. Existing literature supports the relevance of educational toys in early child development, especially smart toys related to digital games that can promote cognitive and motor development in children. These toys not only entertain but also facilitate learning through interaction and personalization. This study validates a prototype of an educational toy using artificial intelligence designed to improve cognitive and motor skills in toddlers aged two to four years, specifically testing on color learning and fine motor skills. It follows an applied research methodology based on the paradigm through design, using participatory mixed methods and design thinking strategies. The validation was carried out in children's educational centers through observational tests and focus groups. The results show that the "Dori" prototype significantly improves hand-eye coordination and facilitates color recognition in children. In addition, it was observed that the toy captures children's attention and promotes effective learning through lights and sounds. The study concludes that intelligent educational toys are valuable tools for the integral development of children, combining playful learning with gaming themes and sensory stimuli. Areas for improvement of this type of toy were also identified, such as the alignment of the pieces and the volume of the sounds, highlighting the importance of integrating technologies in educational toys to promote integral development in children. Nevertheless, it is suggested that they continue investigating their application in diverse cultural contexts and expand the age range of the test participants.

Keywords: play material · artificial intelligence · serious games · children's learning · prototype

1 Introduction

A child's world at an early age is full of discoveries and learning, where each experience contributes to the infant's development. Therefore, their initial education is paramount, focusing on laying the foundations of their future personality, including cognitive, affective, and emotional components, relating them to their personal development [1]. In this

context, as indicated by Prado and Farias [2], play activities have become extremely relevant in children's early education, promoting their creativity, imagination, and integral development. Within these activities, educational toys stand out; in some cases, they are intelligent toys that seek to entertain, educate, and develop children's skills through interactivity and personalization. Many toys are designed for didactic, affective, and value content learning. Especially in the last decade, gamification has become a fundamental tool in children's mental and physical development [3].

However, researchers such as Loredó [4] argue that many parents consider games and toys only as entertainment or amusement, without considering the learning opportunity they represent, even though there are premises that "a child who does not play does not learn." It is also considered that playing with toys is important in children's development, helping them understand the present and prepare for the future, to the point that many artists or scientists attribute their interests to the toys of their childhood [5]. Play is considered the first promoter of cognitive, motor, sensory, and bodily development, inspired by imagination and creativity [6]. Authors and pedagogues agree that toys are essential tools for the good development of the child, representing an important role in the formation of concepts, aptitudes, expectations, and socialization [4], always considering that according to age, the types of games are modified [5]. For example, in children under one year old, games about finding objects hidden in front of their sight, reaching an object, or throwing objects are used; children from one to two years old can start to use mechanical toys to look for objects in various places; and from two years old onwards it is suggested to introduce toys with symbolic games [7].

One type of knowledge children acquire through playing with certain toys is colors, which identify their relationship with other objects and the environment. Distinguishing and recognizing colors among themselves is crucial to making sense of the environment in which they live, acquiring great importance since it allows them to express moods, distinguish one product from another, and improve cognitive reasoning and language [8]. In several cases, technologies have been incorporated into educational toys, radically transforming play dynamics, expecting to achieve learning and improve children's interaction with toys. According to the study by Wang et al. [9], electronic toys for children aged two to four years can be classified based on their characteristics and functionalities. This classification ranges from toys with touch screens that offer interactive play experiences to intelligent educational toys (IET) that stimulate creativity, physical activity, and problem-solving, and in some cases, also related to game application themes. Integrating technology in these toys adds a fun component and intuitively enhances preschool children's fine motor and cognitive development [9].

These toys can interact with children more smartly than traditional toys, offering multimedia features to support personalized learning, progress monitoring, and responses to external stimuli. Therefore, within a didactic context, when considering children two to four years old, it is fundamental to choose technologies that stimulate the senses using multimedia tools [10]. For example, toys that offer interactive and visual activities can be effective for this age group since they encourage exploration and learning through experimentation. However, no studies were found that determine how to establish the validity and use related to designing and implementing an IET.

Thus, considering the background stated, the objective of this research is to validate the use of an IET through the process of designing a functional prototype based on thematic contents for the development of cognitive skills, such as color learning, and the improvement of motor skills, such as fine motor skills, in toddlers between two and four years old. It highlights the importance of using smart electronic toys in the educational context and how they can complement and enrich children's learning experiences at an early age, leading this study to the following research question: How to validate a prototype of an IET using gaming thematic contents, designed to improve cognitive and motor skills in toddlers aged two to four years?

As the foundation for the consolidation of literary antecedents and duly supported conceptual definitions, which must be explained in the context of this study, a theoretical framework is determined in the following subsections.

1.1 Technologies in Smart Educational Toys

Toys have evolved with each historical era. With the Industrial Revolution, toys evolved, extending their manufacture and mechanical objects' antecedents of technological toys [10]. Technological toys include technologies or electronic components to enhance the child's play experience with the toy, which are also identified according to their technological characteristics as smart toys.

A smart toy is an object designed for entertainment and education. Typically, such toys incorporate advanced technology, such as sensors, Internet connectivity, and often artificial intelligence, with ergonomic and aesthetic considerations to enhance the play experience. In these products, the importance of technology as a support for children's education through toys lies in its intimate relationship with learning, capturing children's attention and stimulating the cognitive area [11].

Within the child-computer interaction, a Tangible User Interface (TUI) has been applied to creating educational toys, allowing active or passive interaction with physical objects and digital information. The TUI allows the child to obtain feedback from the actions performed with the toy, indicating, for example, correct actions, time management, or suggestions for use [12]. A technological incorporation that allows the child to receive feedback is light, either natural or artificial. As Yu and Roque [13] point out, embedding lights in a toy enriches the interaction by showing how the child completes a sequence of steps. Lights can function as an educational tool to stimulate children's curiosity and creativity while developing fine motor skills and visual coordination [14].

1.2 The Colors in Early Learning

Colors play a fundamental role in children's early learning, influencing their mood, concentration, and comprehension. The first elements of learning that children encounter are colors, both in objects and in their relationship to their environment, giving meaning to the environment in which they live [8]. The ecological valence theory of Palmer and Schloss [15] suggests that experiences with specific color targets may influence color preferences in children. In addition, color preference is linked to language development and the acquisition of color terms. Zemach and Teller [16] observed that infants are genetically predisposed to certain colors from early on. According to Pitchford,

Davis, and Scerif [17], associating colors with objects facilitates understanding and communication about colors in infants aged two to four. In the educational setting, it is crucial to consider how color affects children's concentration, creativity, and emotional well-being. Light wavelengths influence infants' color preferences, underscoring the importance of designing educational environments that consider the psychological and cognitive impact of color. Color is also fundamental for expressing moods, enhancing motor skills, distinguishing products, and improving cognitive reasoning and language [8].

1.3 Fine Motor Skills

Through play, children explore their identity, make exciting discoveries about their environment, expand their vocabulary, establish social bonds, and experience the healing power of shared laughter. Research also shows that play strengthens the neural connections in children's brains and influences their development on all levels: cognitive, physical, and emotional. During play, children acquire a wide range of complex and adaptive skills. Ann Pleshette Murphy [18] mentions that children aged two to four should reach milestones ranging from playing alone and improving memory to fantasizing or playing pranks with other children. These milestones also allow skills development, such as fine and gross motor skills, which play a fundamental role in this process. Fine motor skills involve precise and coordinated hand and finger movements, which are important for a child's intellectual development [19]. This skill can be developed for activities such as cutting with scissors, completing puzzles, and performing personal care tasks. On the other hand, gross motor skills, which involve larger and more coordinated movements of the whole body, represent an essential part of growth, autonomy, and development in their environment [20] and are fundamental for running, jumping, throwing a ball, and exploring the outdoors. Both types of motor skills allow children to participate in play, thus promoting an integral development in all aspects: cognitive, physical, and emotional.

As mentioned by Cabrera et al. [1]: "They are the movements of the hand and fingers in a precise manner, for the execution of an action with a useful sense, where sight or touch facilitate the location of objects and instruments and is given in almost all actions performed by human beings." Fine motor skills are also intertwined with other areas of child development. For example, the hand-eye coordination needed to perform activities such as drawing or stringing beads on a string helps improve manual dexterity, promoting creativity and artistic expression. In addition, activities that involve manipulating small objects, such as threading needles or building blocks, encourage sensory expression and tactile experimentation, which contributes to a child's sensory and perceptual development.

According to Paz Enrique and Hernández [21], Piaget's constructivism as an approach to educational psychology considers that children learn and construct their knowledge based on the relationship of the school-family-community triad so that students and teachers play an active role in the construction of knowledge, which enhances motivation and participation in the learning process. Furthermore, learning is not the direct result of environmental interactions but is influenced by previously acquired knowledge, i.e., children will learn from iteration with the toy. They will not necessarily get it right the

first time. For Heljakka [22], the value of toys does not lie in their intrinsic value but in the meaning, time, and effort invested in them; it is important to captivate the child's attention to prolong the use of the toy.

The study by Franchin et al. [23] concluded that younger children (up to three years old) need more positive interaction with an adult to have fun with toys. In comparison, older children (six years old) have fun with less interactive behaviors, concentrating more on the toy.

1.4 Design Thinking Applied to the Design of Intelligent Educational Toys

Successful toy design for children goes beyond functionality and aesthetics; it involves a deep understanding of how infants interact with and derive meaning from toys. Therefore, toy designers must strive to align their design decisions with the child's preferences, behaviors, and emotional responses to ensure that the final product resonates with the target audience [22]. Toy design should focus on creating an engaging experience for children; however, some manufacturers prioritize parental preferences to maximize economic gains [24], hence the importance of defining children as users within the Design Thinking process. According to Kudrowitz and Wallace [25], the designer's perception of the play value of a toy concept may differ from the child's final judgment.

The degree of satisfaction must be measured through the product's usability to determine whether the toy developed by applying Design Thinking strategies suits the child. According to ISO 9241-11 standards, usability refers to the fact that designated users (in the case of this research, children from two to four years old) can employ a system, product, or service to achieve predetermined objectives with different degrees of effectiveness, efficiency, and satisfaction in a defined usage scenario.

Usability is evaluated with the help of prototypes that are part of developing the Design Thinking methodology. Usability is influenced by several factors, such as the product itself, users, tasks, and environments, which ultimately affect effectiveness, efficiency, and satisfaction during task execution [26]. Thus, Seyda [27] indicates that Design Thinking is a collaborative, interdisciplinary, innovative, creative, problem-solving, and human-centered process to create user-oriented products, services, or experiences. In addition to analyzing the tastes, preferences, and behavior of children, it is essential to have the criteria of educational psychologists, kindergarten teachers, and the design group, which in turn may be composed of various disciplines such as product design, electronic engineering, communication technologies, robotics, among others, all with different perspectives on the same problem and who contribute their expertise in the design of a toy that meets what the child needs to learn.

2 Methodology

The current applied research, using the paradigm through design, mainly employs mixed methods of a participatory nature. The investigation involves practical experiments seeking to obtain specific data for the validation of the design of a product with a playful and technological approach. It is relevant for this type of research to establish the necessary structure for its development process so that its questions, objectives, and methods are

directly focused on obtaining results concerning the defined problem [28]. For the design and development of the product, prior to the validation of the prototype, Design Thinking strategies were applied as a user-centered methodology that allows the generation of innovative ideas and the exploration of solutions that satisfy the user's needs [29].

This process is divided into five phases (Fig. 1). In the validation phase, a qualitative-quantitative approach was adopted to define the potential use of the IET within the children's educational institutions selected as part of the sample.

In the *empathizing* stage, an immersion process was conducted in an educational environment for children aged two to four years to understand first-hand the educational practices and daily challenges they face. These practices included unstructured interviews with educators from Child Development Centers (CDI), the review of specialized sources, and consultation with the curriculum of the Ministry of Education of Ecuador. Next, a detailed user profile was created to understand the characteristics, interests, and context of the child to whom the IET is intended. In the *Define* stage, the information gathered was consolidated to accurately identify the central challenge of the product to be solved.



Fig. 1. Phases of Design Thinking Applied to the design of an IET

Children's needs were organized by levels, using Maslow's Pyramid as a reference. In addition, a description of the IET, its target audience, characteristics, and purpose were elaborated. Creative solutions were generated through mood boards, brainstorming, and initial sketches in the ideation stage. In addition, the guidance of a psychologist was sought to ensure that the IET is appropriately designed for the user's age. In the *Prototype* stage, physical and digital prototypes of the IET were created for testing and feedback. For the *Validate* phase, concept tests were conducted with children to evaluate if they could complete the toy's mechanics by training their fine motor skills and observation sheets to determine if they identified the basic colors of the IET.

2.1 Validation: Concept Tests and Focus Group

For the prototype design validation, an observational process considered focus group strategies, and a selected sample size from the population established was determined. The segment of the population considered included the participation of caregivers and children from the Child Development Centers (CDI Spanish acronym) administered by the Ministry of Economic and Social Inclusion (MIES Spanish acronym) of Ecuador, with whom an agreement through the Vicente Rocafuerte Higher Technological Institute (ISTVR Spanish Acronym) was obtained for their participation in this research. CDIs

are spaces open to the community whose main objective is to promote the growth and development of children in a situation of social vulnerability. In total, MIES administers 39 CDIs nationwide where, in each CDI, there are around 30 to 50 children between the ages of 12 and 36 months [30], which gives an approximate total population of 1560 children. Particularly, three CDIs in Guayaquil, Samborondon, and Tarifa were more enthusiastic to participate in the focus group. A private day care center related to the researchers' institution also collaborates with this research. From this population, 35 children, distributed in different focus groups or observed individually, were sampled to record and analyze the performance through concept tests with the final prototype developed. The intention was to demonstrate that using this type of toy within the child's educational environment can become a tool to improve fine motor skills and learning of colors during the children's stay in the CDI.

Observations were made using a card that included two groups of variables:

- Fine motor variables: It was observed that the child grasped the pieces firmly, manipulated them correctly with his/her fingers, aligned them before placing them, and adapted easily to changes in their orientation.
- Color learning variables: For this group of variables, it was observed that the child ordered the pieces according to the colors if he/she fit the yellow piece correctly and in the same way as the red and blue pieces.

The data obtained from these observations were tabulated and represented in statistical graphs with measures of central tendency.

In addition, after performing all the concept tests with children, a focus group was conducted with the caregivers who attended the tests to develop a plenary with professional comments on the use and application of the IET. This focus group was conducted online with five caregivers from the different centers where the developed prototype was tested.

3 Results

Following the methodology proposed using Design Thinking for the creation of an IET, the results are presented for each phase determined:

3.1 Empathize Phase

During the unstructured interviews with educators from the CDIs, key activities were identified for children from two to four years of age, such as nesting or puzzles for the development of fine motor skills, while for gross motor skills, games such as “plans-plans” or wooden toys were mentioned, as well as preferences for toys with lights and sounds. The importance of including textile material in the prototype for a more complete interaction was also highlighted (see Fig. 2).

As for the review of specialized sources, it was determined that, for children aged two to four years, play is the main driver of cognitive, motor, sensory, and bodily development by fostering imagination and creativity [6]. In addition, integrating technology into toys captures children's attention and promotes learning through trial-and-error exploration in

play. Likewise, using lights and music in toys improves visual coordination and provides a comprehensive learning experience, combining visual, tactile, and auditory stimuli for complete development [23].

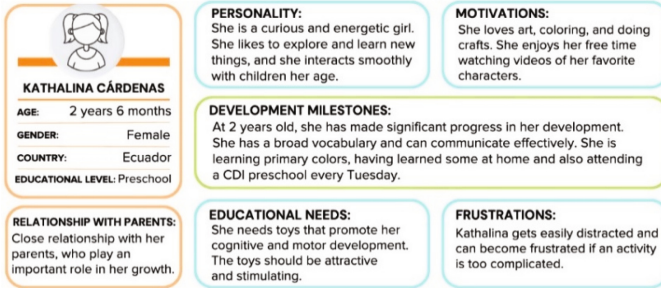


Fig. 2. User profile generated in the empathizing phase

The Ministry of Education’s curriculum contains specific objectives and skills for this age group, highlighting sensory exploration and cognitive development through the identification of colors, shapes, and sounds; this information was used to prepare a brief (Table 1), allowing to understand the context in which the product will be developed.

Table 1. Briefing generated

¿What?	Intelligent Educational Toy (IET)
¿Who?	Aimed at children from two to four years of age
¿How?	Using a toy with motor, sensory, and educational activities, incorporating lights and sound
¿Where?	Children’s Home, Child Development Centers (CDI), Childcare Centers
¿Why?	To foster and provide better learning by developing and strengthening communication, creativity, sensory, and motor skills

3.2 Define Phase

Once the needs of children aged two to four years were identified in the previous phase, the findings were classified using Maslow’s pyramid to illustrate the importance of children’s needs in terms of the potential use of the toy (see Fig. 3). Considering the needs listed on the pyramid, it is defined that the developed IET should promote interaction with the child’s caregiver by using pieces of appropriate size and safe materials for the user’s manipulation. It should also reward the child when completing the interaction, motivating the child to use the IET. Finally, it should encourage creativity as the child develops fine motor skills and learns to identify basic colors.



Fig. 3. Maslow's pyramid generated from children's needs.

3.3 Ideation Phase

In this process phase, the toy's functionality, architecture, and shape were conceived to meet all the requirements of the previous phases. At the beginning of this phase, a conceptual framework was developed (see Fig. 4), which allowed us to define the most important aspects to be considered for the IET design. The researchers highlight the children's use of ergonomics and anthropometry, design trends, shapes, colors, toy regulations, MIDI-am [31] characters, the most suitable materials, and essential electronic components.

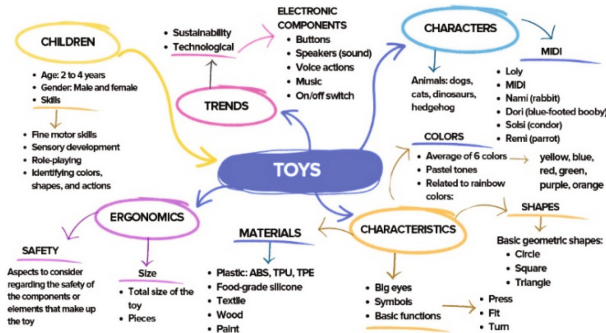


Fig. 4. Conceptual framework developed prior to the design stage

The mood board was used to illustrate inspirations for the ideas and brainstorm with initial sketches to illustrate the ideas generated. (See Fig. 5).

The mood board included analogous products, market trends, and consumer preferences. Additionally, toy design's predominant shapes, colors, and textures were reviewed as a comparative parameter.

At this stage, alternative concepts were also explored, considering world trends in toy design and the characters of the MIDI project [32] so that the design proposals would meet the objective of belonging to the aesthetic line of IET and digital games of a macro project which can compete with global designs.



Fig. 5. Mood board and initial sketches in search of the shape of the toy

3.3.1 Characteristics of the Proposed Design

Based on the information gathered, the IET was designed and named “Dori,” it is about a character from the MIDI-AM projects. From an aesthetic point of view, the character represents a blue-footed booby from the Galapagos Islands. The pieces are rounded with a harmonious and safe appearance for the user. The colors used in the toy’s body are also inspired by the characteristic colors of this species (white, black, and light blue), as seen in the following photograph (see Fig. 6).



Fig. 6. Blue-footed Booby. Photo: Ruben Heleno, CDF, 2011

Regarding conceptual aspects, as shown in the 3D visual prototype (Fig. 7), “Dori” is a character turned into an IET for children from two to four years old, consisting of a base with push buttons and three pieces that must be inserted according to the color. Once each piece is inserted, a light of the color of the push button lights up, and the name of the color is heard, for example, “blue” and so on, until the body of “Dori” is completed. When the child inserts the head, its last piece, an unreleased song from one of the MIDI-am projects, is played.

Additionally, “Dori” features voice playback. A microphone is located on one leg and a speaker on the other, allowing voice interaction with the child. With this activity, the representation of bright colors, basic geometric shapes, and lights helps improve visual perception. At the same time, the repetition of sounds and songs contributes to auditory development and pattern recognition.

“Dori” was designed as an early stimulation tool. As a stackable toy, it improves children’s hand-eye coordination and fine motor skills while manipulating and inserting the pieces. At the same time, it encourages creativity as children experiment with combinations of shapes and colors, and even the incorporation of lights and sounds is a playful element that stimulates the imagination.

Its design is friendly since all the pieces of the IET are safe. For its design, characteristics and measurements of the anthropometry and ergonomics of children from two to four years of age were considered, reducing the risk of suffocation. Safety and resistance were also considered so that the materials and components were durable and non-toxic. In contrast, the materials of the body were translucent to allow observation of the colored lights placed inside each piece. Padded fabric was used for wings, legs, crests, and beaks.

This IET has characteristics of adaptability and versatility. It can modify each body piece’s images, text, or textures, making a visual-sensory connection and thus encouraging creativity.

Dori considers educational elements, using colors that the user identifies and allowing the development of cognitive and sensory skills based on the Ministry of Education curriculum, such as recognition of primary colors, association of basic shapes, and recognition and association of animals through shape.



Fig. 7. 3D visual prototype of “Dori”

Regarding its technical aspects, this IET has an Arduino Nano microcomputer inside that allows the lights and corresponding sound to turn on if the piece is successfully placed. It also has push buttons and LED strips (Fig. 7 shows the toy’s buttons and LED strips). The microcomputer is powered via a USB connection.

The pushbuttons are connected from one end to a 5 V pin of the microcomputer, while the other is connected to its specific pin and a 1000 Ω resistor with output to ground. On the other hand, each LED strip is plugged into its specific pin from the anode, and the cathode goes to the ground. Finally, to play the song at the end of the interaction, an MP3-DFPlayer Mini module for Arduino was used. This circuit is located inside the base of the toy. The buttons and lights are located inside Dori’s pole. A representation of the inclusion of this circuit inside Dori can be seen in Fig. 8.

3.4 Prototyping Phase

Prototypes allow the evaluation of different product characteristics, functionalities, and aesthetics, and there is an infinity of methods, materials, and techniques to carry them

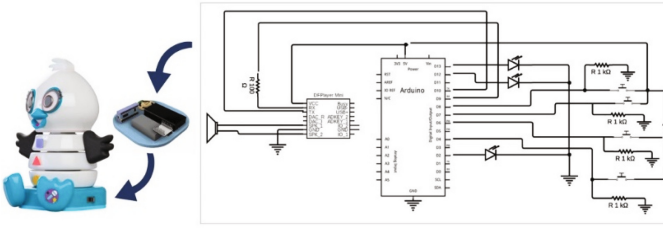


Fig. 8. Designed circuit

out. Different prototypes can be chosen depending on the objectives, resources, and time available, among which we can find digital and tangible or physical prototypes. For this research, in its first phase, several digital prototypes were made in 3D models elaborated in the Autodesk Fusion 360 programs (see Fig. 9), allowing the design capture prior to the elaboration of the physical prototypes. The 3D models allowed the detailed design of the product using parametric modeling tools, section views, exploded views, simulations, photorealistic representations, and animations.



Fig. 9. 3D model designed using Autodesk Fusion 360

For the manufacture of the physical prototypes of the “Dori” IET, we mainly used the method of 3D printing by PLA filament deposition with the help of two printers available in the laboratories of the Faculty of Design and Audiovisual Communication (FADCOM), with two prototypes at different stages of the design process:

- The first low-resolution 3D-printed prototype did not include the technological elements inside, and its objective was to validate the general concept of the IET with children. In this laboratory test, opportunities for improvement were identified, such as size modifications, adjustments to the shape and geometry of the pieces, and the way the pieces fit together.
- The second prototype integrated the electronic components defined in the ideation phase, along with improvements and modifications identified in the prototype.

The second prototype met the needs defined in the ideation phase, in addition to the improvements suggested in the laboratory tests of the prototype, guaranteeing its correct

use, manipulation, and teaching to the child. The final prototype developed can be seen in a photograph in Fig. 10.



Fig. 10. Final prototype developed

3.5 Validation Phase

In this phase, concept tests were carried out with the functional prototype toy developed with children whose parents signed a letter of consent authorizing them to play with their children and take pictures during the session (see Fig. 11). At the end of the tests, a discussion was held with the children's caregivers to receive valuable feedback on the IET and its possible use in the CDIs and Childcare Centers.



Fig. 11. Validation of the final prototype "Dori" (Children's photos are supported by a signed letter of consent from parents)

3.5.1 Proof of Concept

The results presented in Fig. 12 were obtained as a comparative horizontal bar graph for the fine motor variables. As can be seen in Fig. 12, the results were mostly positive. The size chosen was adequate for the children, and they could hold firmly the three pieces and the head of the IET. Likewise, the size and material of the toy favored the manipulation of its parts, except for small problems presented by the children when they placed the blue piece, and the wings of the character got caught between the pieces. Fortunately, the children could correct this problem and place the piece correctly. The greatest problems were encountered in the alignment of the pieces. As can be seen, 43% of the children needed to align the pieces before placing them on the IET base; they

directly placed them without considering a previous alignment that they could do to complete the challenge more easily. However, this was fine since all the children could complete the game. Finally, the children presented slight problems with the changes in the position of the pieces. Some of them placed the pieces in the wrong orientation, which caused the caregiver in charge to correct them. Once this was done, they could quickly place the piece correctly. Favorable results were obtained from synthesizing the data of the motor variables. The IET's construction and pieces allowed the children no major problems completing the challenge.

The variables related to color learning, shown in Fig. 13, were obtained in a comparative horizontal bar graph. Despite this, the children do not, in all cases, consider the order of the colors on the base of the IET when ordering the colors. However, 66% of them achieved this objective.

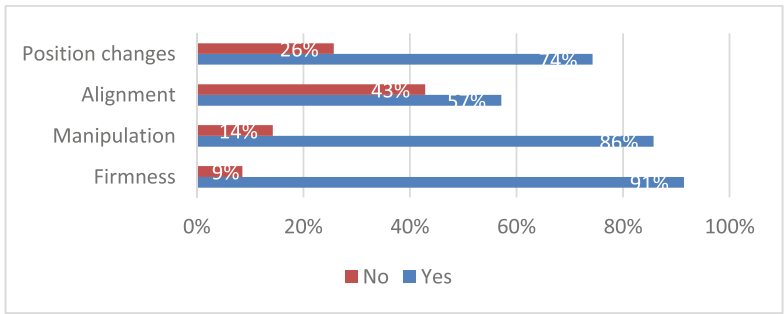


Fig. 12. Results obtained for fine motor variables

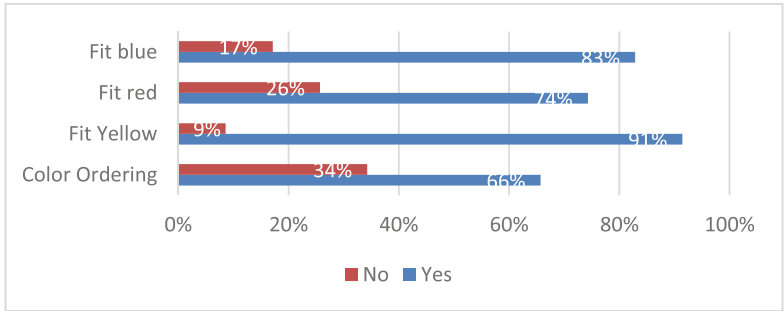


Fig. 13. Results obtained for color learning variables

3.5.2 Users Experience Results - Focus Group

The focus group conducted with the caregivers was held to develop concept testing at the centers with the children. They gave valuable comments regarding the IET aspects and its use. Some examples of their feedback are presented in Table 2.

Table 2. Comments gathered from the focus group.

Category	Registration Unit
Color Learning	The toy works positively for children to learn colors
	Repetitive use of the toy promotes color learning
	The toy's voice and lights attract children's attention and promote learning
Fine Motor Skills	The toy promotes hand-eye coordination and a firm grip
Usability for Different Ages	For children from two to four years old, the version of the toy is enough to use only with one hand
	The toy version requires both hands to fit the pieces together for children from one to two years old

4 Discussion

The study verified that children's learning arises from the interaction with the IET and the repetition of activities, as mentioned by Paz Enrique and Hernández [21]. It is shown how children improved their motor skills and learning of colors thanks to several interactions. It also allowed us to reaffirm that learning is favored thanks to the collaboration and joint participation of children with their caregivers, as argued by Franchin et al. [23]. In addition, it was possible to verify that the interrelated actions between hands and eyes improve the precision in each attempt, as mentioned by Cabrera et al. [1]. On the other hand, it is verified how the use of light as a sensory source encourages children and allows them to enrich their experience with the IET, as is similarly indicated by Yu and Roque [13], in addition to providing the child with feedback on each task completed using the combination of visual, tactile and auditory-musical stimuli [14] deliver a comprehensive learning experience.

Then, valuable comments were collected from the caregivers within the focus group. They indicate that the IET works effectively for children to learn colors. However, that happens more frequently when the use is repetitive. They asserted that regular use of the IET can promote color learning. They indicated that this learning is favored by the toy's voice and the colored lights since they effectively attract children's attention. On the other hand, they confirmed that fine motor skills are favored with the IET, both in hand-eye coordination and firmness of grip.

As for proposed improvements to the IET, the caregivers indicated that the noise in the CDIs is remarkably high, so it would be necessary to increase the volume of the toy's voice, which can be poor sometimes. Another comment was to improve the post where the pieces are placed, as some children found it hard to insert and fit them. Finally, they recommended creating new versions of the IET where it is sufficient to use only one hand to fit the pieces in the case of children from two to four years old, while another version in size exclusively for children from one to two years old where they must use their two hands.

The project also allowed us to verify that collaborative and interdisciplinary work in the IET design process, where the integration of various actors in the subject, was

crucial to obtain better results and mainly the acceptance of the user (the child), as suggested by Seyda [27] regarding the Design Thinking Methodology. The researchers uphold that thanks to the quantitative results of the focus group in the validation phase, the responses were markedly positive for both the variables of fine motor skills and for learning colors. Finally, the IET, with all its functionalities, like the MIDI digital games, is framed according to the objectives and skills proposed in the curriculum required by the Ministry of Education of Ecuador for children of this age. Nevertheless, the design and development of products is an activity susceptible to incremental improvements, and the present research is no exception.

5 Conclusion and Future Work

It is concluded that strong literary evidence examines the relevance of IET in early childhood development, highlighting how these smart toys can promote cognitive and motor development in children. Such toys, whose design process has been validated in this research, entertain and facilitate learning through interaction and personalization. In particular, the functional prototype of IET designed and developed in this study has proven to be an effective tool for improving fine motor skills and color learning in children from two to four years of age. The methodology used in research paradigms for design and methodological strategies such as Design Thinking supported designing and creating an IET prototype that combines technological elements to capture children's attention and playfully facilitate learning.

Concept testing and interviews with educators revealed that the sounds emitted by the IET are a key factor in maintaining children's interest. Children older than three years tended to manipulate the toy with one hand, while two-year-old toddlers used both hands, indicating different levels of motor development. However, areas for improvement were identified, such as the snapping mechanism of the pieces, which tend to get stuck, and the volume of the sounds, which needs to be adjusted. These findings underscore the importance of integrating technologies such as artificial intelligence and other resources such as gaming thematic contents in designing and developing educational toys to foster integrated learning that encompasses cognitive, motor, and sensory aspects to improve cognitive and motor skills in toddlers aged two to four years. Thus, answering the stated research question, we identify some of the most relevant aspects in the design process of this type of toy, the integration of technologies, the pedagogical aspects related to academic content, and predominantly the intervention of the stakeholders involved. These are the children, caregivers, psychologists, and design groups, with whom, through methodological processes focused on the user, such as Design Thinking, facilitate personalizing their interests and needs, especially for toddlers at the center of the learning process. Nevertheless, some suggested design improvements are crucial to increase the effectiveness and attractiveness of the IET prototype. For example, there is a need to adjust the volume and refine the snapping mechanism.

5.1 Future Work

Future research could explore the application of this IET prototype in different cultural contexts and with a wider age range to evaluate its effectiveness in a broader spectrum of

child populations. Incorporating more advanced technologies and customizing the IET to children's individual preferences are also promising areas for development. Besides, it is suggested that more validations be carried out on other CDIs and child care centers to confirm the results for future versions of IET Dori and the creation of similar children's toys.

Overall, this study allows for generating future directions for research, such as testing the toy in different cultural contexts, considering the richness of ethnicities in Ecuador, and broadening the age range of the participants. Additionally, new studies can be developed with other vulnerable groups with physical and psychomotor disabilities for the evaluation of skills.

References

1. de la Caridad Cabrera Valdés, B., de las Nieves Dupeyrón García, M.: El desarrollo de la motricidad fina en los niños y niñas del grado preescolar. *Mendive. Revista Educación* **17**(2), 222–239 (2019)
2. Prado, R.S.A., de Farias, W.S.: Contributions of toys and games in early childhood education. Seven Editora (2023). Accessed: 17 Apr 2024. [Online]. Available: <http://sevenpublicacoes.com.br/index.php/editora/article/view/3243>
3. Vansadiya, R.P., Gondaliya, P.R., Vasoya, N.H., Gupta, S.M.: Recommendations on how educational toys can help preschoolers to improve their social skills. *Educ. Adm. Theory Pract.* **30**(3), Art. No. 3 (2024). <https://doi.org/10.53555/kuey.v30i3.1422>
4. Loredó Abdalá, A.: El juego y los juguetes: un derecho olvidado de los niños. *Acta Pediátrica México* **26**(4), 214–221 (2019). Accessed: 19 Apr 2024. [Online]. Available: <http://repositorio.pediatrica.gob.mx:8180/handle/20.500.12103/1773>
5. Goldsten, J.: El valor de los juguetes y el juego, vol. 1, p. 14 (1997)
6. Ángel Valencia, D., et al.: El juego y los juguetes como aproximación a la alfabetización en educación infantil (2023). Accessed: 22 Apr 2024. [Online]. Available: <https://dspace.tdea.edu.co/handle/tdea/4913>
7. Mozos Pernias, A., López García, M., Pecci Garrido, M.C., Herrero Olaizola, T.: El juego infantil y su metodología, 1st edn, vol. 1 (2010). [Online]. Available: https://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=8934
8. Pérez, A.E.O., Gurieva, N.: Diseño y desarrollo de un libro interactivo para niños: aprender los colores jugando. *Jóvenes Ciencia* **3**(2), Art. No. 2 (2017)
9. Wang, Y., et al.: Exploring the affordances of digital toys for young children's active play. In: Proceedings of the 34th Australian Conference on Human-Computer Interaction, OzCHI'22. Association for Computing Machinery, New York, Apr 2023, pp. 325–337. <https://doi.org/10.1145/3572921.3572935>
10. Magaña, T.: Juguetes que unen el placer con el aprendizaje. *Pinion Education*. Accessed: 26 May 2024. [Online]. Available: <https://pinion.education/es/blog/juguetes-tecnologicos-habilidades-del-futuro/>
11. Pruna, F.X.J., Redín, E.S.E., Mariño, P.A.A., Guarochico, W.F.A.: Los juguetes electrónicos especializados y su influencia en el desarrollo cognitivo social y emocional de los niños entre 5 y 7 años. *RHS-Revista Humanismo Sociedad* **11**(1), Art. No. 1 (2023). <https://doi.org/10.22209/rhs.v11n1a03>
12. Alvarado, M.C., Sanz, C., Baldassarri, S.: Desarrollo de una experiencia educativa con juguetes activos en una mesa de interacción tangible. *Revista de la Asociación Interacción Persona Ordenador (AIPO)* **2**(2), Art. No. 2 (2021)

13. Yu, J., Roque, R.: A review of computational toys and kits for young children. *Int. J. Child-Comput. Inter.* **21**, 17–36 (2019). <https://doi.org/10.1016/j.ijcci.2019.04.001>
14. Trazando Senderos Brillantes: La Importancia del Juego con la Luz en Educación Infantil—San Juan Ikastetxea. Accessed: 30 May 2024. [Online]. Available: <https://sanjuanikastetxea.net/trazando-senderos-brillantes-la-importancia-del-juego-con-la-luz-en-educacion-infantil/>
15. Palmer, S.E., Schloss, K.B.: An ecological valence theory of human color preference. *Proc. Natl. Acad. Sci. U. S. A.* **107**(19), 8877–8882 (2010). <https://doi.org/10.1073/pnas.0906172107>
16. Zemach, I., Chang, S., Teller, D.Y.: Infant color vision: prediction of infants' spontaneous color preferences. *Vision. Res.* **47**(10), 1368–1381 (2007). <https://doi.org/10.1016/j.visres.2006.09.024>
17. Pitchford, N.J., Davis, E.E., Scerif, G.: Does color preference have a role in color term acquisition? *Br. J. Dev. Psychol.* **27**(4), 993–1012 (2009). <https://doi.org/10.1348/026151008X399916>
18. Pleshette Murphy, A.: *El secreto del juego*. Imaginarium (2011)
19. Peñaranda-Peñaranda, M.M., Bayona-Trigos, R.A., Manzano-Durán, O.: Las tecnologías de la información y la comunicación: una herramienta estratégica para el desarrollo académico y profesoral. *Revista Ingenio* **14**(1), Art. No. 1 (2017). <https://doi.org/10.22463/2011642X.2198>
20. Materiales didácticos para el desarrollo de la motricidad gruesa en niñas y niños de 4 años. *Sinergia Académica* (2024). Accessed: 15 May 2024. [Online]. Available: <http://sinergiaacademica.com/index.php/sa/article/view/207>
21. Paz Enrique, L., Hernández, M.: El constructivismo como enfoque en la psicología educativa. *Revista Digital Internacional Psicología Ciencia Social* **8** (2022). <https://doi.org/10.22402/j.rdiipycs.unam.e.8.2.2022.489>
22. Heljakka, K.I.: Toying with creative design: starting to design instruments of play. In: *DS 66-2: Proceedings of the 1st International Conference on Design Creativity (ICDC 2010)* (2010). Accessed: 29 May 2024. [Online]. Available: <https://www.designsociety.org/publication/30279/Toying+With+Creative+Design%3A+Starting+to+Design+Instruments+of+Play>
23. Franchin, L., et al.: Interaction of children with toys to be assembled: a way to promote the development of cognitive and manual skills. *Ergon. Open J.* **4**, 55–60 (2011). <https://doi.org/10.2174/1875934301104010055>
24. Naufal, A., Suzianti, A.: Design thinking approach for product innovation design of educational toys. Presented at the ICEMC 2019: Proceedings of the 2019 5th International Conference on E-business and Mobile Commerce, May 2019, pp. 69–74. <https://doi.org/10.1145/3332324.3332337>
25. Kudrowitz, B., Wallace, D.: The play pyramid: a play classification and ideation tool for toy design. *Int. J. Arts Technol.* **3** (2010). <https://doi.org/10.1504/IJART.2010.030492>
26. Bevan, N.: Measuring usability as quality of use. *Software Qual. J.* **4**(2), 115–130 (1995). <https://doi.org/10.1007/BF00402715>
27. Eraslan Taşpınar, Ş.: Design thinking and art education tasarim odakli düşünme ve sanat eğitimi. *Anadolu Üniversitesi Sanat Tasarım Dergisi* **12**, 379–398 (2022). <https://doi.org/10.20488/sanattasarim.1221700>
28. Puértolas, R.P.: Investigación aplicada en diseño: etapas de la actividad. grafica 93–100 (2024). <https://doi.org/10.5565/rev/grafica.282>
29. Razzouk, R., Shute, V.: What is design thinking and why is it important? *Rev. Educ. Res.* **82**(3), 330–348 (2012). <https://doi.org/10.3102/0034654312457429>
30. Solórzano Alcívar, N., Sornoza Quijije, L.I., Morocho Yunga, G., Pita Quito, G., Poveda Páez, R.: Feasibility of using serious MIDI-AM videogames as resources in early childhood education. In: Maldonado-Mahauad, J., Herrera-Tapia, J., Zambrano-Martínez, J.L., Berrezueta,

- S. (eds.) Information and Communication Technologies. Communications in Computer and Information Science, pp. 429–447. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-45438-7_29
31. Alcivar, N.I.S., Rios, E.S.E., Gallego, D.A.C., Kim, D.H.P., Quijije, L.I.S.: MIDI-AM model to identify a methodology for the creation of innovative educational digital games: a proposed serious game methodology based on university research experiences. In: Improving University Reputation Through Academic Digital Branding, pp. 133–167. IGI Global (2021). <https://doi.org/10.4018/978-1-7998-4930-8.ch009>
 32. Alcivar, N.I.S., Quijije, L.I.S., Gallego, D.A.C.: Evaluations of software developed as research tools to analyze trends of educational videogames characters. *Int. J. Adv. Sci. Eng. Inf. Technol.* **11**(3), Art. No. 3 (2021). <https://doi.org/10.18517/ijaseit.11.3.12431>



Changes in Academic Assessment Due to the Use of Artificial Intelligence

Isaac Ojeda¹ and Santiago Castro Arias²

Universidad Politécnica Salesiana, Cuenca 010105, Ecuador
scastroa1@ups.edu.ec

Abstract. This study seeks to identify effective evaluation methods according to the use of artificial intelligence. It addresses the extent to which artificial intelligence tools and functionalities are used in secondary education, mainly in students and teachers. The scientific literature on the use of artificial intelligence in academic assessment has been reviewed within the methodological framework. In addition, a survey was conducted with students and teachers to gather information on the influence of artificial intelligence, its ethical use, and its application in different academic activities. It is concluded that artificial intelligence is used in various educational activities, allows for the optimization of the work of teachers, and evaluates, improves, and personalizes the teaching-learning process of students. Notably, students who use artificial intelligence tools prefer to be assessed by human teachers. However, these resources should be used as an aid and adequate follow-up. Finally, this work proposes methods of practical evaluation using artificial intelligence.

Keywords: Artificial intelligence · Assessment · Education · Methods secondary · Learning tools

1 Introduction

With the abrupt changes that technology has had and the appearance of artificial intelligence, we find how this is introduced in the educational field. This relationship has caused teachers to modify their teaching strategies and methods since these tools help build resources for each student and improve the teaching-learning process [28]. However, these types of tools propose significant adjustments in terms of ethical issues [6, 7, 19], which produces an essential reflection on the role of teachers in an educational community [6, 13].

The research explores the intervention of artificial intelligence (AI) in assessment, mainly in secondary education. It starts with an analysis of the literature that helps to understand that artificial intelligence tools have transformed and will transform academic evaluation, and then shows the view of students and teachers on its use. This research helps teachers search for strategies based on

Supported by Universidad Politécnica Salesiana.

artificial intelligence that contribute to the development of the teaching-learning process and, at the same time, allows them to analyze their limitations and challenges.

It is exposed that the research on Artificial Intelligence (AI) in education is broad and diverse, ranging from theoretical studies to practical applications [28,30]. Research reveals that AI has been used to personalize learning and improve outcomes by applying more accurate and tailored assessments [6,7,19]. However, discussion persists about the reliance that can be placed on these types of tools and their impact on educators.

The purpose of this paper is to establish effective evaluative methods for the use of artificial intelligence in secondary education by exploring the literature and students' and educators' perceptions of its use. From this analysis, the study seeks to answer the question: What are the effective evaluative methods for the use of artificial intelligence in secondary education?

2 Related Works

Artificial intelligence (AI) has been incorporated into education, transforming the role of teachers and opening up new possibilities for learning. However, its use and implementation bring with it ethical issues that must be carefully considered [6,7,19].

AI is the process by which machines perform tasks, think, learn, and act, just as humans do, without replacing humans [16]. The incorporation of these types of tools in the educational environment has invited to rethink the role of teachers, forcing them to train and work in an integrated way with these types of tools, which help them to understand the educational development of students, to correct tasks and evaluations, to anticipate the academic performance of students and also in the development of school planning [1,3,10,30].

Collaboration between the educational community and industry positively aids pedagogical processes [22]. While artificial intelligence is a fundamental pillar in any area of knowledge and industry, one should maintain sight of the dangers that may arise from its use in education. Among them are the influence of technologies in academia, the weakening of essential academic-pedagogical activities and the diminishing role of educators in evaluation [13]. In the face of these dangers, it is crucial to address the responsibility for using AI as an educational tool; this responsibility lies with teachers and academic institutions, who cannot ignore the existence of AI, underestimate its importance, or exaggerate its effect [30].

Research on academic assessment and AI has highlighted countries such as China, the United States, and the United Kingdom between 2019 and 2023 (Fig. 1). Universities with the most research in this field are also mainly located in China (Fig. 2).

Artificial Intelligence in education is an underexplored area, mainly in combining educational strategies to create meaningful learning experiences for students, as expressed by [27,29]. When combined with Artificial Intelligence and

other tools, educational assessment can increase its ability to support the teacher. An example is STEM assessment, which can benefit from AI in performing laboratory tasks and obtaining a clear assessment [7]. Information presentation is evolving toward virtual and mixed reality, facilitating assessment by combining AI with the physical part [6, 19]. On the other hand, videos become an essential tool for learning and assessment, driving innovation in education; in the same vein, in China, AI has reduced inequality in the distribution of educational resources [18]. In the same way, the development of immersive modules, the implementation of digital pedagogy, and the use of virtual technologies for education open new possibilities for assessment. Collaborative work becomes fundamental to creating strategies that can be complemented with a technological model that mitigates this limitation, although they do not allow an individualized analysis of each student [26]. Educational institutions need to include AI in education through adaptive learning; for this purpose, it is proposed to establish teaching-learning methods with the help of intelligent devices developed with AI; among them, we can mention virtual reality, cloud computing and portable information as expressed by [28]. The scientific literature reviewed about AI applied to education and mainly in evaluation analyzes and identifies topics related to the present research as presented in Table 1 below. Different issues are shown, among them adaptive learning, which, as its name indicates, refers to the search for and implementation of strategies that meet the individual needs of each student. Similarly, the topic of anti-plagiarism contains the use of artificial intelligence tools to detect plagiarism in written work. In addition, the evaluation instruments that allow for estimating students' academic performance are discussed. The methods and methodologies of artificial intelligence evaluation are also analyzed, showing the different ways of evaluating within the teaching-learning process. Finally, references are made to using artificial intelligence in education to improve this process. The studies present the use of artificial intelligence tools based on large language models in formal student assessments, these researches raise important considerations to take into account about the very being of academic assessment because while they can support the education of students including the teaching of writing and composition among others, they limit the ability of students in the development of critical thinking skills, research and the embodiment of original ideas creation of original and coherent texts and can evade existing technological methods and trained academic staff.

3 Problem Formulation and Methodology

The evolution of technology, especially Artificial Intelligence (AI) in the educational field, still generates doubts about its implementation. However, it's important to note that AI, when effectively integrated, has the potential to enhance the teaching-learning process. Research has been conducted on artificial intelligence in education, but there is still a lack of knowledge about its identification and implementation. This lack of knowledge causes concern about its proper use, to the point of thinking that they could replace the role of teachers.

Table 1. Evaluation methods in education and the application of artificial intelligence.

Author	Adaptive learning	Anti-plagiarism written by AI	Analytical rubrics	Classification algorithms	AI evaluation methods and methodologies	AI in education
Perkins 2023 [23]	–	–	X	X	X	X
Elkhatat 2023 [11]	–	X	–	–	X	X
Niu et al. 2022 [21]	X	–	–	X	X	X
Rahm and Rahm-Skågeby 2023 [24]	–	–	–	–	X	X
Fergus, Botha, and Ostovar 2023 [12]	–	X	–	–	X	X
Bütüner and Calp 2022 [4]	–	–	–	–	X	X
Kaldaras, Yoshida, and Handek 2022 [17]	–	–	X	X	X	X
Harati et al. 2021 [15]	X	–	–	–	X	X
Halimi and Seridi-Bouchelaghem 2021 [14]	–	–	–	X	X	X
Chaudhry et al. 2023 [5]	–	X	–	–	X	X
Aimeleweh 2022 [2]	X	–	–	–	X	X
Crawford, Cowling, and Allen 2023 [8]	–	X	–	–	X	X
Crompton and Burke 2023 [9]	–	–	–	–	–	X
Seo et al. 2021 [25]	–	–	–	–	–	X
Mangera, Supratno, and Suyatno 2023 [20]	–	–	–	X	X	X
Present work	X	X	X	X	X	X

In addition to this need for more knowledge and concerns about the use of artificial intelligence, there are also the challenges involved in evaluating these tools in the educational environment, mainly due to the application students can give to them. Although the implementation still generates doubts and questions, its capacity to transform education is essential.

To address this set of concerns, the present research proposes to explore how teachers and students can effectively and efficiently integrate AI in secondary education by identifying assessment methods that are compatible with the use of AI. Thus, the research will provide input for future research on adapting these tools to the educational environment.

3.1 Methodology

It begins with a review of the literature on artificial intelligence in education, followed by the application of surveys to teachers and students, and finally, the data obtained were analyzed. Three stages are proposed for this purpose.

Stage 1: Review of scientific literature. Using the historical-descriptive methodological approach, scientific articles have been explored in recognized databases such as Web of Science, Springer and Scopus. The keywords used for the search are: “academic evaluation”, “artificial intelligence”, “secondary education”, and “evaluation”. They selected articles published between 2021 and 2023, which yielded 401 results. Discrimination by areas of research knowledge was made, focusing on education, computer science, science and technology. Finally, with the help of VosViewer, a bibliometric analysis was performed, and the countries and universities with the most research on this topic were identified.

Stage 2: Survey application. Two surveys were designed; one applied to 66 teachers and the other to 124 high school students. Different questions were asked, such as demographic data, knowledge about academic assessment with AI support, the use of this tool in education, the frequency of its use, the application and their view of the relationship between artificial intelligence and humans.

For teachers, the survey consists of 39 questions. It was answered by teachers from both private and public educational institutions, from 18 to 55 years

of age, with experience in teaching from elementary school to high school, in different areas of knowledge of the Ecuadorian national educational curriculum and from other regions of the country. The students answered 14 questions, ages between 12 and 18 years old, from different Ecuador regions and public or private educational institutions. As shown in Tables 2, 3 and 4.

Table 2. Survey applied to high school and high school students.

Questions	Answers
1. Age:	12-13 years — 14-15 years — 16-17 years — 18 years and over
2. Gender identity:	Male — Female — Non-binary
3. What grade are you in?	Eighth grade — Ninth grade — Tenth grade — First year of Bachillerato — Second year of Bachillerato — Third year of Bachillerato
4. In which region is your educational institution located?	Coast — Highlands — Amazon — Island
5. What device do you use most frequently to study?	Mobile — Tablet — Computer — All equally
6. What type of support does your educational institution have?	Public — Private — Fiscomisional — Municipal
7. Have you heard about artificial intelligence?	a. No, I've never heard of it — b. I've heard of artificial intelligence, but I don't know much about it. — c. I am moderately familiar with it. — d. Yes, I am aware of what it is and how it works.
8. What kind of school work do you do with the help of artificial intelligence?	a. Searching for information on the internet. — b. Automatic translation of texts. — c. Correcting grammar and spelling. — d. Summarising text. — e. Solving mathematical problems. — f. Solving physics problems. — g. Speech recognition for text transcription. — h. Creating slides. — i. Video creation. — j. Music creation. — k. Creation of social media content. — l. Other:
9. Select the subjects in which you have used artificial intelligence for your school work.	a. Math. — b. Physics. — c. Chemistry. — d. Biology. — e. Language and literature. — f. Social studies. — g. English. — h. Artistic education. — i. Computing. — j. Physical education. — k. STEAM. — l. Programming and database. — m. History. — n. Philosophy. — o. Citizenship education. — p. Entrepreneurship and management. — q. Technical support. — r. Programming. — s. Operating systems. — t. Office automation applications. — u. Marketing. — v. Other: Google Translate. — DeepL. — Grammarly. — LanguageTool. — Resoomer. — TLDR. — Siri. — Alexa. — Duolingo. — Khan Academy. — Kahoot! — Classcraft. — ChatGPT. — Bard - Gemini. — Jasper. — Notion. — Canva. — Lensa. — Cymath. — Smodin. — MuseNet. — Buffer.
10. Of the following artificial intelligence tools, which ones have you used?	
11. Do you consider it ethically correct to carry out a school assignment using exclusively artificial intelligence?	Yes. — No. — Maybe
12. What are your main concerns about the use of artificial intelligence to do school homework?	Loss of own learning skills. — Academic honesty. — Quality of results. — Dependence on technology. — I have no concerns. — Other:
13. If you had the choice, would you prefer to have your homework and exams graded by artificial intelligence or by a human teacher?	Artificial Intelligence. — Human Teacher. — Both. — Neither.
14. Why would you choose the former option? - Please provide a brief explanation.	

Stage 3: Analysis and synthesis of information. Finally, using the analytical-synthetic method, with the help of Microsoft Power BI and Public, the results of the surveys are analyzed. Convergences and divergences between the different questions and using Artificial Intelligence are identified. Finally, after each result, an analysis of the results obtained is performed.

Table 3. Survey applied to high school and high school teachers.

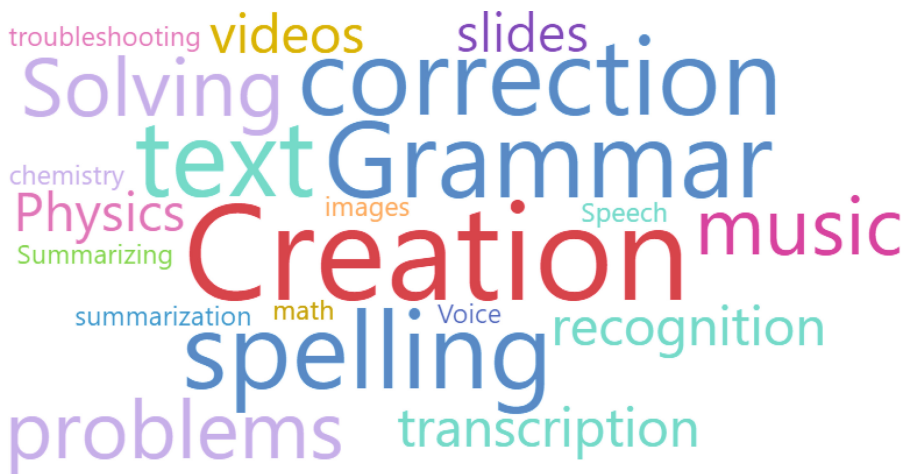
Questions	Answers
1. Age.	18–25 years. — 26–35 years. — 36–45 years. — 46–55 years. — Over 55 years old.
2. Gender identity.	Male — Female — Non-binary — Prefer not to answer
3. Level of education.	Baccalaureate. — Third level. — Fourth level.
4. Level(s) at which you teach?	a. Upper basic education. — b. Baccalaureate.
5. Which area(s) do you teach?	Cultural and Artistic Education. — Physical Education. — Natural Sciences. — Social Sciences. — Language and Literature. — Mathematics. — Foreign Language. — Other:
6. How many years have you been teaching?	0–5 years. — 6–10 years. — 11–15 years. — 16–20 years. — More than 20 years.
7. In which region is the educational institution where you work located?	Coast. — Highlands. — Amazon. — Island.
8. What type of support does the educational institution where you work have?	Public. — Private. — Fiscomisional. — Municipal.
9. Which device do you use most frequently to prepare your assessments?	Mobile. — Tablet. — Computer. — All equally.
10. In general, how would you rate your level of skill with technology?	High. — Medium. — Low
11. Have you heard of Artificial Intelligence in educational assessment?	Yes. — No.
12. How often do you use Artificial Intelligence tools in your teaching?	Daily. — Several times a week. — Weekly. — Monthly. — Hardly ever. — Never.
13. To what extent are you aware of the different applications of Artificial Intelligence in educational assessment?	Very little. — Regular. — Quite a lot. — A lot.
14. how do you usually find out about new Artificial Intelligence tools applied to educational assessment?	Seminars or workshops. — Online courses. — Education blogs. — Recommendations from other teachers — Other:
15. Which Artificial Intelligence tools have you used for plagiarism detection?	Turnitin. — Copyscape. — Grammarly. — Not used. — Other:
16. Which Artificial Intelligence tools have you used to create rubrics?	Rubistar. — Classcraft. — Google Classroom. — Not used. — Other:
17. Which Artificial Intelligence tools have you used to classify results?	Gradescope. — Khan Academy. — Not used. — Other:
18. Which Artificial Intelligence tools have you used to implement evaluation methods and methodologies?	Khan Academy. — DreamBox Learning. — Kahoot! — Classcraft. — Google Arts — Culture. — Metaverse. — Not used. — Other:
19. For which evaluation tasks have you used Artificial Intelligence tools?	Grading of exams. — Correction of assignments. — Monitoring of student progress. — Personalisation of assessment. — Not used. — Other:
20. What is your level of confidence in the decisions made by Artificial Intelligence tools in educational assessment?	Scale of 1 to 5, where 1 is “little confidence” and 5 is “a lot of confidence”.
21. What do you think is the main obstacle to implementing Artificial Intelligence in educational assessment in your institution?	Lack of staff training. — Lack of financial resources. — Resistance to change. — Lack of support. — Other:

4 Analysis of Results

Figure 3 shows that higher primary and high school students use artificial intelligence (AI) tools in various school tasks. The applications with more use are content creation with 50% of respondents, grammar and spelling with 40%, and solving with 20% of respondents. AI is used to a lesser extent for problem-solving, music and videos, with 10%. The above shows that students use artificial intelligence to support their learning and improve academic performance in areas of knowledge such as writing, comprehension and research.

Table 4. Survey applied to high school and high school teachers.

Questions	Answers
22. What factors do you consider important when choosing an AI tool for educational assessment?	Ease of use. — Accuracy in grading. — Adaptability to students' needs. — Cost. — Other:
23. In your opinion, what are the main advantages of using artificial intelligence in educational assessment?	Increased efficiency in assessment. — Increased grading accuracy. — More accurate detection of plagiarism. — Personalisation of assessment. — Increased fairness in assessment. — Other:
24. Research on the integration of artificial intelligence in teaching practice.	Development of materials and resources for teacher training. — Designing strategies for the implementation of artificial intelligence in the classroom. — Creation of learning communities to share experiences. — Other:
25. Research on the development of new artificial intelligence tools for evaluation.	Exploration of new technologies such as machine learning and natural language processing. — Designing tools that are adaptable to different educational contexts. — Developing tools that are easy to use for teachers and students. — Other:
26. Would you be interested in being trained to use artificial intelligence tools in educational assessment?	Definitely. — Probably. — Not sure. — Probably not. — Definitely not.

**Fig. 3.** School tasks performed with the help of artificial intelligence.

Next, the areas of knowledge in which the surveyed students have used Artificial Intelligence for their school activities are identified. Seventy-three per cent of the students surveyed mentioned that programming is used for tasks related to the subject. Next, 71% are language and literature, and 68% of the respondents. Biology is in fourth place with 59%, followed by chemistry with 49%. This shows a trend towards integrating artificial intelligence in various academic tasks to improve learning and productivity. This integration especially benefits activities related to computer science; however, the use of AI in other areas of knowledge shows a growing interest in applying these tools. As can be seen in Fig. 4.

Students were asked about their use of Artificial Intelligence tools. Through the above word cloud analysis, it is observed that the most used Artificial Intel-



Fig. 4. Areas of knowledge in which artificial intelligence has been used for school tasks.

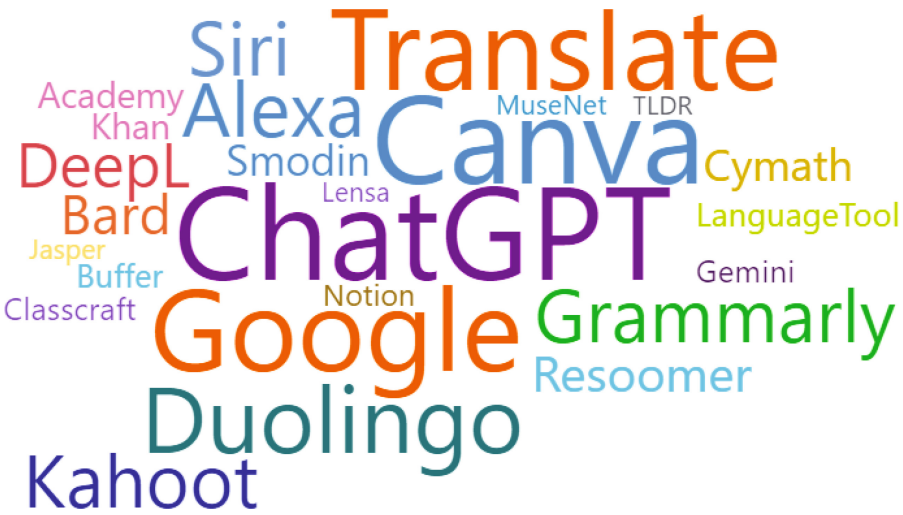


Fig. 5. Use of artificial intelligence tools by students.

lignce tool is Chat GPT, with 80% of students surveyed choosing it, followed by Canva, with a choice of 75%, Translate and Google 69% of students surveyed, Duolingo 53%, Kahoot and Grammarly 39%, Google Gemini with 69, among others.

Figure 5 shows the significant use of artificial intelligence tools, highlighting students' preferences. This helps teachers understand their needs and improve the educational resources they propose.

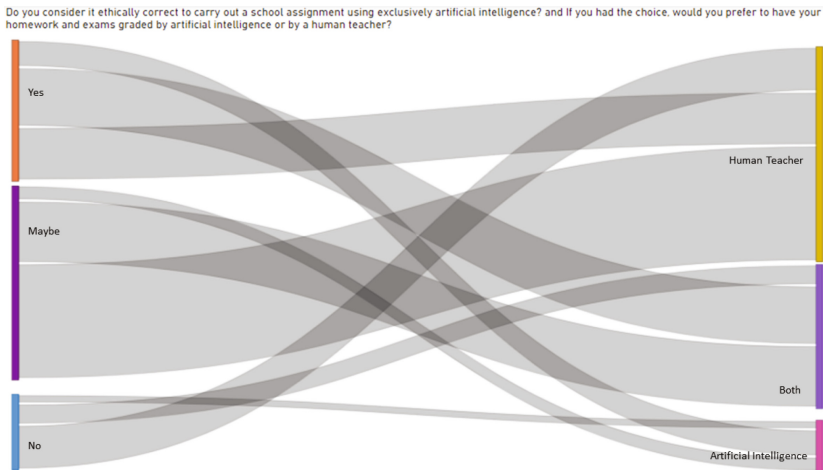


Fig. 6. Ethics of using only AI in school assignments and the preference between grading by a human teacher or AI.

The analysis presented in Fig. 6 involves two critical questions. The first question asks whether respondents consider it ethical to perform a school assignment using only Artificial Intelligence, while the second question asks whether they would prefer to have their assignments and assessments graded by a human teacher or by Artificial Intelligence.

As a result of this cross-analysis, we obtained 15% of respondents who consider using Artificial Intelligence to perform tasks ethically but prefer a human teacher to grade their assignments and assessments. On the other hand, 30% of students responded “maybe” to the first question and chose a human to grade their homework and exams. Likewise, 10% of students responded to the survey saying that it is unethical to do assignments with AI alone; the same students report that they also prefer humans to grade their assignments and exams. It is essential to note that 13% of students responded in the survey that it is ethical to do assignments with AI only; they also prefer to be graded by a human teacher. Still, 15% of respondents answered that “maybe” it is ethical to do assignments with AI only, and they prefer to be graded by humans and by AI.

The above shows the complexity of students’ attitudes towards using artificial intelligence in their academic activities. A group considers the exclusive use of artificial intelligence in school tasks. In contrast, another group expresses ethical concerns and shows a preference for human intervention in the grading of assessments and assignments. Therefore, it is considered essential to address not only students’ technological capabilities but also ethical and pedagogical considerations when integrating artificial intelligence into the teaching-learning process.

On the one hand, students use AI tools; on the other hand, they have concerns about using them for different school tasks, as shown in Fig. 7. The concern that

the use of AI produces the loss of personal skills stands out in 80% of students, as does the academic issue combined with honesty in 80%. Consequently, although students use various AI tools, there are specific concerns about their application in education. As previously observed, there is a wide use of artificial intelligence tools by students; however, this adoption of technology and academic integrity produces significant concerns among students regarding the loss of personal skills due to this dependence on artificial intelligence, as well as the concern about being detected by tools of this same line, which can lead to risks due to misuse.

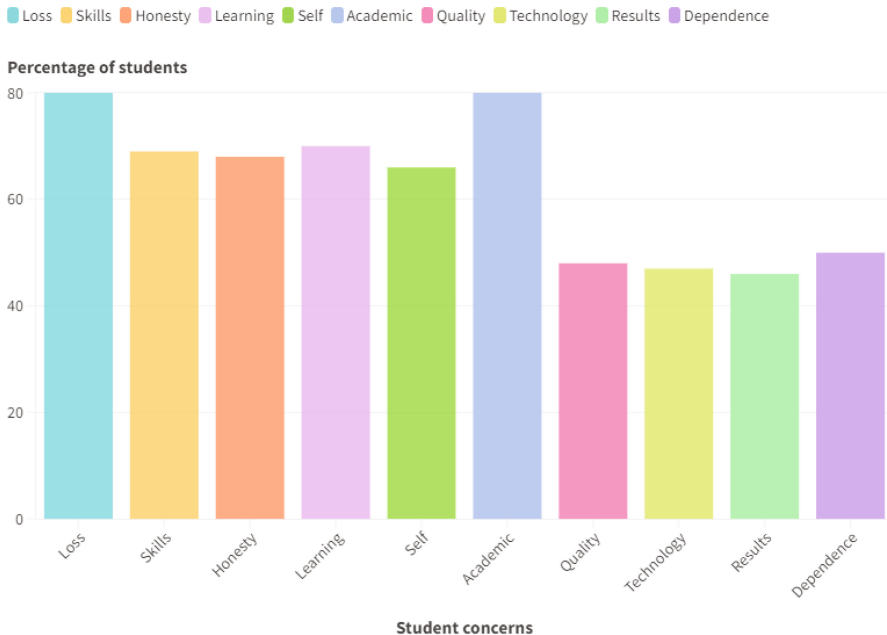


Fig. 7. Students' concerns about the use of AI in performing tasks.

Two questions in the survey addressed two important aspects related to the area in which teachers carry out their activities and the frequency of implementation of AI tools. As shown in Fig. 8, 100% of mathematics, computer science, language, and literature teachers use AI tools daily, while 75% in foreign language and 60% in natural science are registered among the respondents. Finally, 20% of cultural and artistic education respondents implement AI in their teaching activities daily. Implementing AI tools in teaching activities presents several significant messages since the teachers of the knowledge areas show high frequency in using these tools.

Figure 9 shows the use of artificial intelligence tools to detect plagiarism. Turnitin is used by 50% of the teachers surveyed in cultural and artistic education, followed by natural sciences 46%, foreign language 45%, language and literature

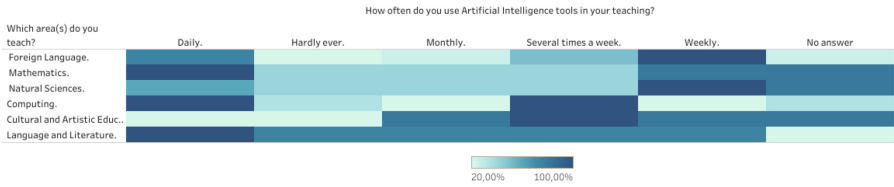


Fig. 8. Areas of knowledge in which they teach and frequency of use of artificial intelligence tools in teaching.

42%, mathematics 40%, and computer science 28%. Grammarly is used mainly by the areas of knowledge of computing and language and literature with 28%, followed by foreign language with 18%, continuously by the area of cultural and artistic education with 14% and mathematics with 10% and natural sciences with 6%. On the other hand, Copyscape is mainly used by teachers of foreign languages with 27%, followed by computer science and language and literature with 14%; in turn, it was found to be used by teachers of the areas of knowledge of natural sciences and cultural and artistic education with 7%. It is also evident that, in the area of mathematics, this tool is not used. The importance of the fact that a significant number of teachers in different areas of knowledge do not use Artificial Intelligence tools for this type of activity is also highlighted. Although an interesting number of teachers surveyed used AI tools for plagiarism detection, further training is needed in the area of originality detection. In addition, Fig. 10 examines teachers’ practices in the areas of knowledge taught about using AI tools to classify results. Thirty per cent of teachers in foreign language and computer science use this tool; 20% of teachers in the knowledge areas of mathematics, natural sciences and cultural and artistic education use it. Only 20% of teachers in mathematics and natural sciences and 10% of foreign language teachers use Gradescope. In this sense, it is essential to emphasize the need for further exploration and adoption of AI tools in educational assessment, considering the different needs and learning contexts in implementing educational technology.

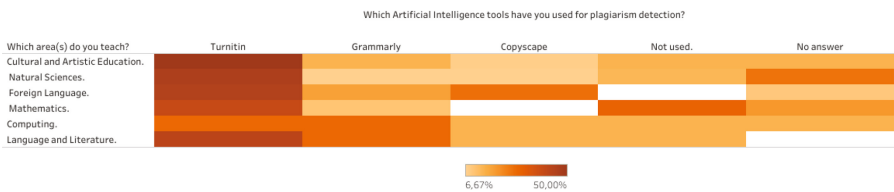


Fig. 9. Use of artificial intelligence tools for plagiarism detection.

Figure 11 shows the use of Artificial Intelligence tools to implement evaluation methods and methodologies, contrasted with the area where the surveyed

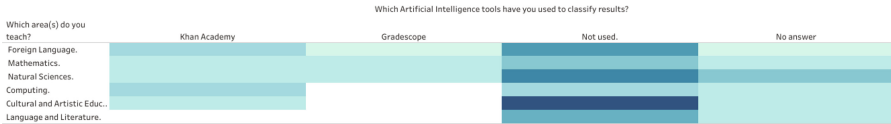


Fig. 10. Use of artificial intelligence tools for results classification by teachers.

teachers teach. The results show that 80% of the teachers in Cultural and Arts Education use Kahoot!, while 70% of the foreign language teachers and 50% of the language and literature teachers also use it. In addition, 40% of mathematics and natural sciences teachers use these tools, followed by 30% of computer science teachers. On the other hand, Khan Academy is used by 50% of natural science teachers, 30% of foreign language teachers, and 20% of mathematics and computer science teachers. Finally, only 10% of cultural and arts education and language and literature teachers use artificial intelligence tools to implement evaluation methods and methodologies. As for teachers who have not used these tools, they represent 30% in Cultural and Artistic Education, 20% in mathematics, natural sciences and computer science, and 10% in language and literature.

These data suggest that, although teachers are increasingly using artificial intelligence tools to implement evaluation methods and methodologies, a significant number of teachers still lack knowledge about the adoption of technological tools of this type in their different areas of knowledge, thus creating opportunities to promote integration with educational evaluation.

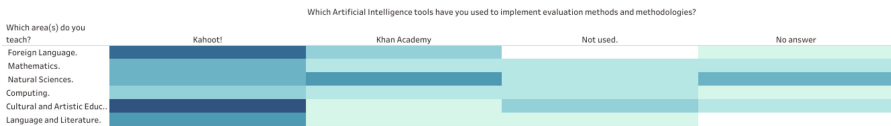


Fig. 11. Use of artificial intelligence tools for the implementation of evaluation methods and methodologies.

Regarding the use of these tools by teachers in their assessment activities about the area in which they teach, Fig. 12 shows that in the area of natural sciences, five teachers use them for monitoring learning, one teacher uses them for creating and correcting assessments and two more use them for customization. However, they were only used for customization with tests, and four participants preferred not to respond to this survey. Four mathematics teachers use AI tools for student progress monitoring. In contrast, one teacher uses it to make and customize assessments, two others do not use AI tools for these purposes, and two prefer to refrain from responding. About five teachers in the foreign language area use AI tools to correct assignments, while two use them to personalize and monitor students' progress. One teacher did not respond to this question, while one indicated not using AI tools. In computing, two teachers use their

AI grading system for personalization. One of the teachers uses it for student progress monitoring, another has yet to use the tools, one prefers not to answer, and one does not use it.

While five teachers in the Cultural and Arts Education area do not use the AI for this type of activity, four use these tools to construct evaluations, one to monitor student progress, and one to correct exams. Finally, two teachers in language and literature use the personalization and grading tools, two do not, and one uses them to monitor students' progress. These results demonstrate a diversity of approaches to integrating AI into educational assessment.

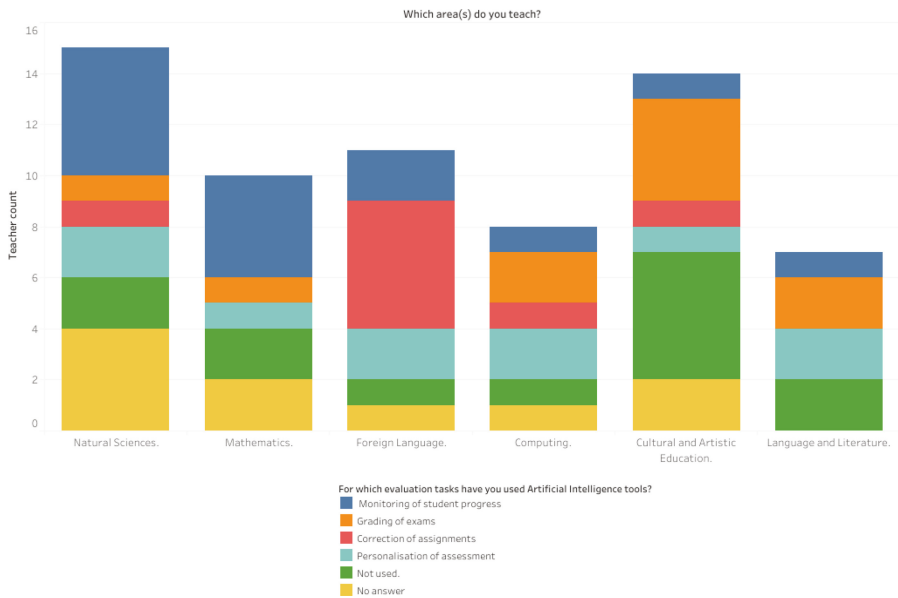


Fig. 12. Evaluation tasks using artificial intelligence tools by teachers.

Finally, Fig. 13 shows the results of the current work that was analyzed. It is established that the effective evaluation methods according to the use of artificial intelligence in secondary education are: The first will lie in adaptive learning as it will guide how to implement learning systems that help us to use Artificial Intelligence to adjust the content and assessments to the educational needs of each student. From this perspective, learning could be personalized, and skills and knowledge development could be assessed more accurately and equitably. Second, the training, implementation and use of anti-plagiarism tools in secondary education. The third method will involve the development of analytical rubrics based on artificial intelligence. The fourth proposed method lies in applying ranking algorithms to student data, ranking students from the highest performers to those requiring more attention and fostering collaborative work

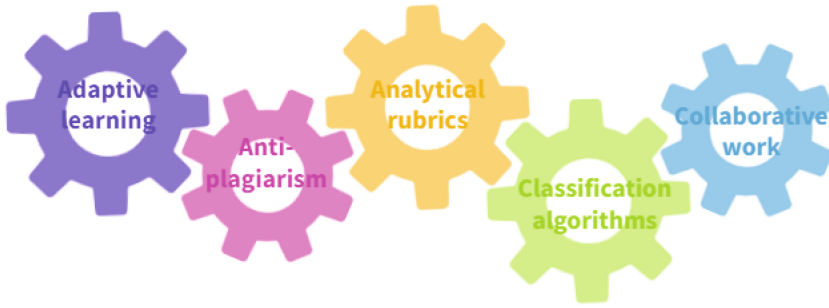


Fig. 13. Effective evaluation methods according to the use of artificial intelligence in secondary education.

with AI-based platforms, which will evaluate each student's final product, collaboration, communication and individual contribution.

5 Conclusions

Artificial Intelligence tools have not just changed, but transformed teaching-learning strategies and methods, opening up a world of possibilities to personalize and optimize this process. They also raise ethical issues and the crucial need to redefine the role of teachers and educators, sparking a new era of education.

Consequently, artificial intelligence tools can be applied to multiple areas of knowledge; specifically, plagiarism detection or the creation of analytical rubrics can extend student assessment to various more objective ways and provide a more in-depth evaluation. In addition, grading algorithms can grade students with different performances in the same way, which is revolutionary for educational assessment.

This work has been important for creating AI-aligned assessment methods that could support future research on the appropriateness of this technology for education. The review of scientific literature and the survey of teachers and students have helped us better understand the impact and appreciation of using AI in academic assessment. High school students' use of AI for various school activities indicates that these technologies are gradually being integrated into the educational process.

This paper highlights the need for teamwork between technology developers, teachers, and students to design and implement artificial intelligence solutions that effectively respond to the learning process needs.

Acknowledgments:. The GITEE research group and the Universidad Politécnica Salesiana have supported this work.

References

1. Ahmad, S.F., Alam, M.M., Rahmat, M.K., Mubarik, M.S., Hyder, S.I.: Academic and administrative role of artificial intelligence in education. *Sustainability (Switzerland)* **14**(3), 1–11 (2022). <https://doi.org/10.3390/su14031101>
2. Almelweth, H.: The effectiveness of a proposed strategy for teaching geography through artificial intelligence applications in developing secondary school students' higher-order thinking skills and achievement. *Pegem Egit. Ogr. Derg.* **12**(3), 169–176 (2022). <https://doi.org/10.47750/pegegog.12.03.18>
3. Baashar, Y., et al.: Evaluation of postgraduate academic performance using artificial intelligence models. *Alex. Eng. J.* **61**(12), 9867–9878 (2022). <https://doi.org/10.1016/j.aej.2022.03.021>
4. Bütüner, R., Calp, M.H.: Estimation of the academic performance of students in distance education using data mining methods. *Int. J. Assess. Tools Educ.* **9**(2), 410–429 (2022). <https://doi.org/10.21449/ijate.904456>
5. Chaudhry, I.S., Sarwary, S.A.M., El Refae, G.A., Chabchoub, H.: Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT – a case study. *Cogent Educ.* **10**(1) (2023). <https://doi.org/10.1080/2331186X.2023.2210461>
6. Chen, S.S.C., Duh, H.: Interface of mixed reality: from the past to the future. *CCF Trans. Pervas. Comput. Interact.* **1**(1), 69–87 (2019). <https://doi.org/10.1007/s42486-018-0002-8>
7. Christopoulos, A., Pellas, N., Laakso, M.J.: A learning analytics theoretical framework for STEM education virtual reality applications. *Educ. Sci.* (2020)
8. Crawford, J., Cowling, M., Allen, K.A.: Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). *J. Univ. Teach. Learn. Pract.* **20**(3) (2023). <https://doi.org/10.53761/1.20.3.02>
9. Crompton, H., Burke, D.: Artificial intelligence in higher education: the state of the field. *Int. J. Educ. Technol. High. Educ.* **20**(1) (2023). <https://doi.org/10.1186/s41239-023-00392-8>
10. Cruz-Jesus, F., et al.: Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon* **6**(6), e04081 (2020). <https://doi.org/10.1016/j.heliyon.2020.e04081>
11. Elkhatat, A.M.: Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *Int. J. Educ. Integr.* **19**(1), 1–23 (2023). <https://doi.org/10.1007/s40979-023-00137-0>
12. Fergus, S., Botha, M., Ostovar, M.: Evaluating academic answers generated using ChatGPT. *J. Chem. Educ.* **100**(4), 1672–1675 (2023). <https://doi.org/10.1021/acs.jchemed.3c00087>
13. Gendron, Y., Andrew, J., Cooper, C.: The perils of artificial intelligence in academic publishing. *Crit. Perspect. Account.* **87**, 102411 (2022). <https://doi.org/10.1016/j.cpa.2021.102411>
14. Halimi, K., Seridi-Bouchelaghem, H.: Students' competencies discovery and assessment using learning analytics and semantic web. *Australas. J. Educ. Technol.* **37**(5), 77–97 (2021). <https://doi.org/10.14742/ajet.7116>
15. Harati, H., Sujo-Montes, L., Tu, C.H., Armfield, S.J., Yen, C.J.: Assessment and learning in knowledge spaces (Aleks) adaptive system impact on students' perception and self-regulated learning skills. *Educ. Sci.* **11**(10) (2021). <https://doi.org/10.3390/educsci11100603>

16. Hervieux, S., Wheatley, A.: Perceptions of artificial intelligence: a survey of academic librarians in Canada and the United States. *J. Acad. Libr.* **47**(1), 102270 (2021). <https://doi.org/10.1016/j.acalib.2020.102270>
17. Kaldaras, L., Yoshida, N.R., Haudek, K.C.: Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Front. Educ.* **7**, 1–15 (2022). <https://doi.org/10.3389/educ.2022.983055>
18. Li, J., Tan, X., Hu, Y.: Research on the framework of intelligent classroom based on artificial intelligence. *Int. J. Electr. Eng. Educ.* (2021). <https://doi.org/10.1177/0020720920984000>
19. Makransky, G., Borre-Gude, S., Mayer, R.E.: Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *J. Comput. Assist. Learn.* **35**(6), 691–707 (2019). <https://doi.org/10.1111/jcal.12375>
20. Magera, E., Supratno, H., Suyatno: Exploring the relationship between transhumanist and artificial intelligence in the education context: particularly teaching and learning process at tertiary education. *Pegem Egit. Ogr. Derg.* **13**(2), 35–44 (2023). <https://doi.org/10.47750/pegegog.13.02.05>
21. Niu, S.J., Luo, J., Niemi, H., Li, X., Lu, Y.: Teachers' and students' views of using an AI-aided educational platform for supporting teaching and learning at Chinese schools. *Educ. Sci.* **12**(12) (2022). <https://doi.org/10.3390/educsci12120858>
22. Pantanowitz, L., et al.: Rules of engagement: promoting academic-industry partnership in the era of digital pathology and artificial intelligence. *Acad. Pathol.* **9**(1), 100026 (2022). <https://doi.org/10.1016/j.acpath.2022.100026>
23. Perkins, M.: Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *J. Univ. Teach. Learn. Pract.* **20**(2) (2023). <https://doi.org/10.53761/1.20.02.07>
24. Rahm, L., Rahm-Skågeby, J.: Imaginaries and problematisations: a heuristic lens in the age of artificial intelligence in education. *Br. J. Educ. Technol.* **54**(5), 1147–1159 (2023). <https://doi.org/10.1111/bjet.13319>
25. Seo, K., Tang, J., Roll, I., Fels, S., Yoon, D.: The impact of artificial intelligence on learner–instructor interaction in online learning. *Int. J. Educ. Technol. High. Educ.* **18**(1) (2021). <https://doi.org/10.1186/s41239-021-00292-9>
26. Sepasgozar, S.M.: Digital twin and web-based virtual gaming technologies for online education: a case of construction management and engineering. *Appl. Sci. (Switzerland)* **10**(13) (2020). <https://doi.org/10.3390/app10134678>
27. Wu, J., Ou, G., Liu, X., Dong, K.: How does academic education background affect top researchers' performance? Evidence from the field of artificial intelligence. *J. Informet.* **16**(2), 101292 (2022). <https://doi.org/10.1016/j.joi.2022.101292>
28. Xie, H., Chu, H.C., Hwang, G.J., Wang, C.C.: Trends and development in technology-enhanced adaptive/personalized learning: a systematic review of journal publications from 2007 to 2017. *Comput. Educ.* **140** (2019). <https://doi.org/10.1016/j.compedu.2019.103599>
29. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int. J. Educ. Technol. High. Educ.* **16**(1) (2019). <https://doi.org/10.1186/s41239-019-0171-0>
30. Zhu, Z.: Composition of online teaching and academic ability under the background of artificial intelligence and HTML. In: *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, pp. 1467–1470 (2021). <https://doi.org/10.1109/ICCMC51019.2021.9418250>

Author Index

A

- Alcivar, Nayeth Idalid Solorzano 345
Alejandro, Ricardo J. 57
Almeida-Galárraga, Diego 39, 100
Alonso-Anguizaca, Jose 117
Ambrosi, Kevin Sebastian Campoverde 131
Armas, Raúl 285
Astudillo, Jaime 71, 145, 253

B

- Ballaz, Santiago J. 178
Banegas, Diego Patricio Peña 193
Basurto-Cruz, Edgar 117
Bermeo, Alexandra 131
Buestán, Néstor Rafael Salinas 193

C

- Cabrera, Danny 160
Cáceres-Bení, Karen 100
Cadena-Morejón, Carolina 39
Calderon-Cordova, Carlos 318
Cárdenas Villacrés, David H. 333
Castillo, Darwin 57
Castillo-Malla, Darwin 86
Castro Arias, Santiago 364
Celi, Patricio Eduardo Ramon 131
Chachalo, Roberth 253
Chamorro, Kevin 178
Charco, Jorge L. 117
Coello-Andrade, Mateo David 3
Cruz-Varela, Jonathan 39

G

- García, Geovanny 303
García, Santiago 57
García-Montero, Patricio Santiago 131, 238, 271
Ghia, Elias 303
Gonzalez, Anthony Xavier Romero 131

- Gualán, Ramiro Israel Vivanco 86
Guevara, Cesar 39

H

- Huilcapi, Víctor 303

I

- Infante, Saba 253

J

- Jiménez-Gaona, Yuliana 86
Jiménez-Ruescas, Javier 71, 145
Juárez-Gonzalez, Bryan 39

K

- Kim, Da Hee Park 345

L

- Lakshminarayanan, Vasudevan 57, 86
León, Juan Pablo Astudillo 178
Lima, Juan Fernando 238
Ludeña-González, Patricia 208
Lupercio, Jonnathan Emmanuel Cubero 238

M

- Marcillo-Vera, Ana 100
Martinez-Mosquera, Diana 160
Masabanda, Rommel 160
Morocho-Cayamcela, Manuel Eugenio 3, 21, 223
Moya, Alejandro 71, 145

N

- Narváez, Ariana Deyaneira Jiménez 223
Navia, Marlon 285

O

- Ojeda, Isaac 364
Orellana, Marcos 131, 238, 271

P

Pineda, Israel 253
Pinos, Pablo Andres Molina 271
Pozo, Jimmy Ernesto Canizares 345
Pretel, Elena 71, 145
Puchaicela, Victor 318

Q

Quijije, Ángel Iván Torres 193

R

Recalde, Lorena 160
Reyes, Iván 21, 223
Reyes-Chacón, Iván 3
Rodríguez, Diego Fernando Intriago 193
Rodríguez-Álvarez, María José 57
Rodríguez-Alvarez, María José 86
Rohoden, Katty 208

S

Salas Ibarra, Raquel de los A. 333
Sánchez, Andrés 208
Sarango, Roger 318
Sierra, Michael Xavier Arce 345
Soto, Brian 303
Suntaxi, Gabriela 160

T

Tirado-Espí, Andrés 100
Tirado-Espín, Andrés 39
Torres Romero, Alexander E. 333
Tualombo, Mayra Elizabeth 21

V

Varela, Francisco Alexander Zambrano 193
Vera, Dánely Leonor Sánchez 223
Villalba-Meneses, Fernando 39
Vizcaíni-Imacaña, Paulina 21
Vizcaíno-Imacaña, Paulina 223
Vizcaino-Imacaña, Paulina 3
Vizcaíno-Imacaña, Paulina 39

Y

Yanza-Montalvan, Angela 117

Z

Zambrano-Martinez, Jorge Luis 131, 238,
271
Zhinin-Vera, Luis 71, 145
Zumba-Gamboa, Johanna 117
Zurita, Andrea Paola Rubio 345