Juan Antonio Lossio-Ventura · Rosario Medina Rodriguez · Jack Zilberman · Freddy Alberto Paz Espinoza · Hugo Alatrista-Salas (Eds.)

Communications in Computer and Information Science 2895

Information Management and Big Data

12th Annual International Conference, SIMBig 2025
Lima, Peru, October 29–31, 2025
Proceedings

Springer

SIMBig25

Communications in Computer and Information Science 2895

Series Editors

Gang Li, *School of Information Technology, Deakin University, Burwood, VIC, Australia*

Joaquim Filipe, *Polytechnic Institute of Setúbal, Setúbal, Portugal*

Zhiwei Xu, *Chinese Academy of Sciences, Beijing, China*

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 http://link.springer.com/bookseries/7899. Conferences publishing in CCIS are allowed to use our online conference service (Meteor) 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

Juan Antonio Lossio-Ventura ·
Rosario Medina Rodriguez · Jack Zilberman ·
Freddy Alberto Paz Espinoza ·
Hugo Alatrista-Salas
Editors

Information Management and Big Data

12th Annual International Conference, SIMBig 2025
Lima, Peru, October 29–31, 2025
Proceedings

Springer

Editors
Juan Antonio Lossio-Ventura
National Institutes of Health (NIH)
Bethesda, MD, USA

Jack Zilberman
Universidad Peruana de Ciencias Aplicadas
Lima, Peru

Hugo Alatrista-Salas
De Vinci Higher Education, De Vinci
Research Center
Paris, France

Rosario Medina Rodriguez
INRIA
Paris, France

Freddy Alberto Paz Espinoza
Pontificia Universidad Católica del Perú
Lima, Peru

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-032-20321-2 ISBN 978-3-032-20322-9 (eBook)
https://doi.org/10.1007/978-3-032-20322-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2026

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

SIMBig 2025[1], the 12th edition of the International Conference on Information Management and Big Data, was held between October 29-31, 2025 at the Universidad Peruana de Ciencias Aplicadas[2], Lima Peru.

SIMBig 2025 introduced innovative approaches for analyzing and handling datasets as well as new methods based on Artificial Intelligence (AI), Data Science, Machine Learning, Natural Language Processing, Semantic Web, Data-Driven Software Engineering, Health Informatics, and more.

SIMBig seeks to promote cooperation between national and international students, researchers, practitioners, and professors to enhance data-driven decision-making by developing and deploying new methods for data analysis. SIMBig provides a welcoming environment where participants present their scientific contributions through full and short papers.

This book contains the proceedings of the papers presented at the 12th edition of SIMBig. For this edition, 36 long papers and 1 short paper were selected for publication from 101 submissions, resulting in an acceptance rate of 36.6%. Submissions received three double-blind reviews each, on average. SIMBig 2025 featured seven keynote speakers who are experts in the field of AI. These keynote speakers came from institutions in Canada, the USA, and France. They covered several topics relevant to the audience, as described below.

Diana Inkpen, from the University of Ottawa, Canada, presented a talk titled "Explainable Neural Text Classification". She discussed how advances in Large Language Models (LLMs) enable the development of highly accurate neural text classifiers, while also exacerbating the challenge of explainability due to their black-box nature. Inkpen described her work on developing explainable neural text classifiers that partially opens this black box. Explainability can be addressed either at the model level or at the level of individual classification decisions. She emphasized that explanations need to reflect both what is learned from training data and what is encoded in the underlying pre-trained LLM. For explaining individual decisions, she discussed feature importance methods such as LIME, SHAP, and Integrated Gradients. She also presented approaches for generating full-text explanations through customized prompting or joint learning of classes and explanations. Inkpen showed results from two case studies in legal text mining and mental health text mining. The generated explanations were evaluated using both automatic metrics and human judgment to assess their relevance and usefulness.

Tina Hernandez-Boussard, from Stanford University, USA, delivered a talk entitled "Patient-Specific Digital Twins for Learning Health Systems". She discussed digital twins—virtual representations of individual patients that integrate multimodal clinical, biological, and environmental data—and their potential to advance precision medicine. Hernandez-Boussard highlighted the critical challenge of ensuring that patient-specific

[1] https://simbig.org/SIMBig2025/

[2] https://www.upc.edu.pe/en/

simulations are accurate, safe, transparent, and ethically governed. She outlined ongoing work on digital twin frameworks that combines causal and mechanistic modeling with real-world EHR data to simulate care trajectories and predict treatment response. She emphasized both the technical foundations and the necessary guardrails for responsible use. Ultimately, she described patient-specific digital twins as a pathway toward continuously learning healthcare systems, where models acted as adaptive partners in improving care quality, safety, and fairness.

Dan Goldstein, from Microsoft, USA, delivered a talk titled "Toward AI Interactions That Build Rather Than Erode Human Cognitive Abilities". He examined common concerns that delegating cognitive work to AI may erode human thinking and learning, while emphasizing that not all AI interactions are alike. Goldstein argued that a more productive question is not whether AI helps or harms learning in general, but under what conditions and through what mechanisms it can improve human capabilities, as measured by performance on subsequent unaided tasks. He presented findings from studies in which participants received assistance from LLMs while solving math problems, drafting professional correspondence, or interpreting legal documents. The LLMs used were lightly modified versions of commercial chatbots rather than purpose-built tutoring systems. Across all tasks, participants who received AI support outperformed control participants on follow-up problems. He proposed three mechanisms that could explain these gains, noting that while AI use can contribute to skill atrophy, it might also facilitate inadvertent learning.

Ravi Kumar, from Google, USA, presented "Algorithmic Problems in Discrete Choice". His talk focused on discrete choice settings, where a user selects one option from a finite set of alternatives, a process central to recommendation systems in e-commerce, social media, and search engines. He discussed Random Utility Models (RUMs), a popular framework in which users assign values to options and select the one with the highest value. RUMs have become increasingly important in the Web era, offering an elegant mathematical framework for modeling and predicting user behavior from limited observations. Although RUMs have been extensively studied in behavioral economics and the social sciences, Kumar highlighted that many fundamental algorithmic tasks remain poorly understood. He discussed several algorithmic and learning questions related to RUMs.

Marc Najork, from Google, USA, presented a talk entitled "Generative Information Retrieval". His presentation provided a brief history of Information Retrieval and Natural Language Understanding, two areas of Computer Science that evolved separately and are now converging to enable a new class of Question Answering systems. Generative Information Retrieval systems answer a user's information need directly, rather than referring the user to relevant sources. These systems draw both on parametric knowledge encoded in the weights of LLMs and on external tools, such as search engines, knowledge bases, or program execution sandboxes. Najork also discussed several challenges ahead, including the impact of such systems on the web ecosystem.

Ariel Procaccia, from Harvard University, USA, presented a talk titled "Thinking Outside the Ballot Box". He addressed the challenge of designing novel democratic processes capable of handling enormous sets of alternatives, such as all possible policies, bills, or statements. Dr. Procaccia argued that this challenge can be addressed through a

framework called generative social choice, which fuses the rigor of social choice theory with the flexibility and power of LLMs. He explored an application of generative social choice to the problem of identifying a proportionally representative slate of opinion statements. This included a discussion of desired properties, an algorithm that provably achieves them, an implementation using GPT, and insights from an end-to-end pilot. By providing guarantees, generative social choice can alleviate concerns about AI-driven democratic innovation and help unlock its potential.

Pierre Zweigenbaum, from CNRS, France, presented "Digital Commons for Generative AI on French Clinical Text: PARTAGES". He described PARTAGES as a large French consortium comprising 10 health NLP research teams, 20 hospitals, and 2 startups, including Mistral, led by the French National Platform for Health Data (the Health Data Hub). The PARTAGES project, funded by BPI France, aims to build and share open-source French resources related to LLMs for health. Due to the sensitivity of medical reports, GDPR-related constraints in Europe make it difficult to disseminate health-specialized LLM resources. To overcome these challenges, the project was designed around open-source deliverables, including a medium-sized French medical LLM; a corpus of 5,000 fictitious medical reports written by specialists with annotations for several use cases; a data augmentation algorithm for medical reports; seven specialized algorithms corresponding to medical use cases; a platform for federated evaluation with 20 nodes across 20 hospitals; and methodological guides to support stakeholders in adapting models to their own contexts or replicating the approach in other languages.

December 2025

Juan Antonio Lossio-Ventura
Rosario Medina-Rodríguez
Jack Zilberman
Freddy Alberto Paz Espinoza
Hugo Alatrista-Salas

Organizing Committee

General Organizers

Juan Antonio Lossio-Ventura	National Institutes of Health, USA
Hugo Alatrista-Salas	De Vinci Higher Education, De Vinci Research Center, Paris, France

DISE Track Organizers

Jorge Eduardo Díaz Suárez	Universidad Peruana de Ciencias Aplicadas, Peru
Eduardo Ceh-Varela	Eastern New Mexico University, USA
Freddy Alberto Paz Espinoza	Pontificia Universidad Católica del Perú, Peru

Local Organizers

Jack Zilberman	Universidad Peruana de Ciencias Aplicadas, Peru
Miguel Nunez-del-Prado	Universidad Peruana de Ciencias Aplicadas, Peru
Patricia Urrea Rodriguez	Universidad Peruana de Ciencias Aplicadas, Peru
Mark Christhian Barrueta Pinto	Universidad Peruana de Ciencias Aplicadas, Peru
Ricardo Moscoso Cuaresma	Universidad Peruana de Ciencias Aplicadas, Peru

Program Co-chairs

Juan Antonio Lossio-Ventura	National Institutes of Health, USA
Rosario Medina-Rodríguez	Inria, France
Jack Zilberman	Universidad Peruana de Ciencias Aplicadas, Peru
Freddy Alberto Paz Espinoza	Pontificia Universidad Católica del Perú, Peru
Hugo Alatrista-Salas	De Vinci Higher Education, De Vinci Research Center, Paris, France

Program Committee

Main Track Program Committee

Abel García Nájera	Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico
Abraham Dávila	Pontificia Universidad Católica del Perú, Peru
Abteen Ebrahimi	University of Colorado Boulder, USA
Adel Taweel	Birzeit University, Palestine
Adrien Coulet	Inria, France
Agn'es Braud	Université de Strasbourg, France
Agnieszka Mykowiecka	Institute of Computer Science, Polish Academy ofSciences, Poland
Ahmed Azough	De Vinci Research Center, France
Aidan Hogan	Universidad de Chile, Chile
Alain Abran	École de Technologie Superieure, University of Quebec, Canada
Alan Valejo	Federal University of São Carlos, Brazil
Alberto Ochoa-Zezzatti	Universidad Autónoma de Ciudad Juárez, Mexico
Aleksandra Karpus	Gdańsk University of Technology, Poland
Alejandro Flores Lamas	Royal Holloway, University of London, UK
Alexandra González-Eras	Universidad Técnica Particular de Loja, Equador
Alexandre Alves	CMCC/UFABC, Brazil
Alexander Rodríguez	University of Michigan, USA
Ali Khenchaf	ENSTA Bretagne, Lab-STICC CNRS UMR 6285, France
Aline Villavicencio	University of Sheffield, UK
Alistair Plum	University of Luxembourg, Luxembourg
Allaoua Chaoui	University Constantine 2, Algeria
Alti Adel	University Ferhat Abbas SETIF-1, Algeria
Amedeo Napoli	LORIA, Nancy, France
Amel Borgi	Tunis El Manar University, Tunisia
Ana-Maria Bucur	Università della Svizzera italiana, Switzerland
André Miralles	INRAE, France
Andreas Speck	Kiel University, Germany
Andrew Gentles	Stanford University, USA
Andrew Tomkins	Google, USA
Aneesh Krishna	Curtin University, Australia
Aneta Poniszewska-Maranda	Lodz University of Technology, Poland
Annika Schoene	Northeastern University, USA
Anisa Rula	University of Brescia, Italy
Antonela Tommasel	ISISTAN, CONICET-UNICEN, Argentina

Antonio Mauttone	Universidad de la República, Uruguay
António Rosado da Cruz	Instituto Politécnico de Viana do Castelo, Portugal
Ari Barrera Animas	Universidad Panamericana, Mexico
Ariel Sabiguero	Universidad de la República, Uruguay
Audrone Lupeikiene	Vilnius University, Lithuania
Aurea Soriano-Vargas	Universidad de Ingeniería y Tecnología, Peru
Bell Manrique Losada	Universidad de Medellín, Colombia
Benet Vincent	Coventry University, UK
Bernd Amann	LIP6, Sorbonne Université, CNRS, France
Billy Peralta	Universidad Andres Bello, Chile
Bogumila Hnatkowska	Wrocław University of Science and Technology, Poland
Brett Drury	Liverpool Hope University, UK
Bruno Martins	Instituto Superior Técnico, University of Lisbon, Portugal
Camilo Thorne	Elsevier, Germany
Carlos Bobed Lisbona	University of Zaragoza, Spain
Carlos Vazquez	École de Technologie Supérieure, Canada
Cédric Grueau	Lisbon School of Engineering, Portugal
Cédric Lopez	Emvista, France
Cengiz Acarturk	Jagiellonian University, Poland
Cesar A. Collazos	Universidad del Cauca, Colombia
Charles Zheng	National Institute of Mental Health, USA
Cesar A. Collazos	Universidad del Cauca, Colombia
Christian Sallaberry	Université de Pau et des Pays de l'Adour, France
Christian von Lucken	Universidad Nacional de Asunción, Paraguay
Christelle Scharff	Pace University, USA
Christophe Debruyne	Université de Liège, Belgium
Christophe Rodrigues	De Vinci Higher Education, De Vinci Research Center, France
Claude Tadonki	Mines Paris - PSL University, France
Claudio de Farias	Universidade Federal do Rio de Janeiro, Brazil
Clemente Rubio-Manzano	Universidad del Bío-Bío, Chile
Cristian López Del Alamo	Universidad de Ingeniería y Tecnología, Peru
Cristina Seceleanu	Mälardalen University, Sweden
Damiano Distante	University of Rome Unitelma Sapienza, Italy
Daniel Wladdimiro	De Vinci Higher Education, De Vinci Research Center, France
Dante Mújica-Vargas	CENIDET, Mexico
David Hicks	Texas A&M University-Kingsville, USA
David La Red Martínez	Universidad Nacional del Nordeste, Argentina
Dayang N. A. Jawawi	Universiti Teknologi Malaysia, Malaysia

Dijana Oreski	University of Zagreb, Croatia
Diana Inkpen	University of Ottawa, Canada
Diana Martinez	Escuela Politécnica Nacional, Equador
Dina Demnar-Fushman	National Library of Medicine, USA
Divine Iloh	University of Arkansas at Little Rock, USA
Dominique Gay	Université de La Réunion, France
Edelweis Rohrer	Universidad de la República, Uruguay
Edgar Ceh Varela	Eastern New Mexico University, USA
Edgar Rios Piedra	Stanford University, USA
Edgar Sarmiento	Universidad Nacional San Agustín, Peru
Edson Melanda	Federal University of São Carlos, Brazil
Edward Hinojosa-Cardenas	Universidad Nacional San Agustín, Peru
Edwin Alvarez Mamani	Universidad Peruana de Ciencias Aplicadas, Peru
Edwin Montes	UAM Cuajimalpa, Mexico
Eddy Sánchez-DelaCruz	Tecnológico Nacional de México, Mexico
Eduardo Blanco	University of Arizona, USA
Elena Beretta	Vrije Universiteit Amsterdam, Netherlands
Eloísa García-Canseco	Universidad Autónoma de Baja California, Mexico
Emma Franklin	Birmingham City University, UK
Engelbert Mephu Nguifo	Université Clermont Auvergne, CNRS, LIMOS, France
Eric Umuhoza	Carnegie Mellon University, USA
Erick Antezana	Norwegian University of Science and Technology, Norway
Ernestina Menasalvas	Universidad Politécnica de Madrid, Spain
Esaú Villatoro-Tello	Idiap Research Institute, Switzerland
Eva Rolin	UCLouvain, Belgium
Evelyn Cervantes	Amazon Web Services, Brazil
Fabio Lopes	Universidade Presbiteriana Mackenzie, Brazil
Facundo Quiroga	Universidad Nacional de La Plata, Argentina
Farah Ait Salaht	De Vinci Higher Education, De Vinci Research Center, France
Félix Castro Espinoza	Autonomous University of the State of Hidalgo, Mexico
Fedra Trujillano	University of Glasgow, UK
Fevzi Belli	University of Paderborn, Germany
Florence Le Ber	Université de Strasbourg / ENGEES, France
Freddy Paz	Pontificia Universidad Católica del Perú, Peru
Gabriel Assis	IC/UFF, Brazil
Gabriel Jimenez	Paris Brain Institute, France
George Kakarontzas	University of Thessaly, Greece

Georgia Koloniari	University of Macedonia, Greece
Gerardo Sierra	Universidad Nacional Autónoma de México, Mexico
Giacomo Bergami	Newcastle University, UK
Giancarlo Sperli	University of Naples Federico II, Italy
Gianluca Quercini	CentraleSupélec, France
Gloria Piedad Gasca-Hurtado	Universidad de Medellín, Colombia
Gordana Rakic	University of Novi Sad, Serbia
Gregor Stiglic	University of Maribor, Slovenia
Guillaume Metzler	University Lumière Lyon 2, Laboratoire ERIC, France
Guillermo Calderon-Ruiz	Universidad Católica de Santa Maria, Peru
Hafedh Mili	Université du Québéc à Montréal, Canada
Hamidreza Vanaei	De Vinci Higher Education, De Vinci Research Center, France
Harvey Siy	University of Nebraska at Omaha, USA
Hiram Calvo	Center for Computing Research - Instituto Politécnico Nacional, Mexico
Huei Diana Lee	Western Paraná State University, Brazil
Hugo Alatrista-Salas	De Vinci Higher Education, De Vinci Research Center, France
Hugo Jair Escalante	INAOE, Mexico
Hugo Estrada-Esquivel	Tecnológico Nacional de México, Mexico
Ian Horrocks	University of Oxford, UK
Imen Ouled Dlala	De Vinci Higher Education, De Vinci Research Center, France
Ismael Osuna-Galán	Universidad Autónoma del Estado de Quintana Roo, Mexico
Issa Atoum	World Islamic Sciences and Education, Jordan
Ivan Dimitry Ribeiro Zyrianoff	University of Bologna, Italy
Jacinto Dávila Quintero	Universidad de Los Andes, Colombia
Jacinto Estima	University of Coimbra, Portugal
Jaroslav Pokorny	Charles University, Czech Republic
Javier Montero	Universidad Complutense de Madrid, Spain
Javier Sanchez-Galan	Universidad Tecnológica de Panamá, Panama
Jesus Bellido	Universidad de Ingeniería y Tecnología, Peru
Jhon Quiza	Universidad de Medellín, Colombia
Jakub Swacha	University of Szczecin, Poland
Jean-Paul Calbimonte	University of Applied Sciences and Arts Western Switzerland HES-SO Switzerland
Jérôme Gensel	Université Grenoble Alpes, France
Jesús Rosa-Bilbao	Universidad de Cádiz, Spain
Jiawei Han	University of Illinois Urbana-Champaign, USA

Jinan Charafeddine	De Vinci Higher Education, De Vinci Research Center, France
Johanna Cordova	INALCO, France
John Ortega	University of Santiago de Compostela, Spain
Jordi Nin	BBVA Data & Analytics, Spain
Jorge Eduardo Díaz Suárez	Universidad Peruana de Ciencias Aplicadas, Peru
Jorge Lopez-Vargas	Universidad Técnica Particular de Loja, Equador
José Alanis	Instituto Nacional de Investigaciones Nucleares, Mexico
José Carlos Ortiz-Bayliss	Tecnológico de Monterrey, Mexico
José Eduardo Ochoa Luna	San Pablo Catholic University, Peru
José Fabián Reyes Román	Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Spain
José Herrera	Universidad Nacional Mayor de San Marcos, Peru
Julio Garcia	University of Calgary, Canada
Juan Antonio Lossio-Ventura	National Institutes of Health, USA
Juan Carlos Olivares Rojas	Instituto Tecnológico de Morelia, Mexico
Juan Gutierrez-Cardenas	Universidad de Lima, Peru
Juan Lazo Lazo	Universidad del Pacífico, Peru
Juan Pablo Bascur Cifuentes	Leiden University, Netherlands
Julian Grigera	Universidad Nacional de La Plata, Argentina
Julio César Mello Román	Universidad Nacional de Asunción, Paraguay
Julio Garcia	University of Calgary, Canada
Julio Ricardo Moscoso Cuaresma	Universidad Peruana de Ciencias Aplicadas, Peru
Jun Kong	North Dakota State University, USA
Ka Chun Lam	National Institute of Mental Health, USA
Kemal Polat	Bolu Abant Izzet Baysal Üniversitesi, Turkey
Konstantin Läufer	Loyola University Chicago, USA
Krung Sinapiromsaran	Chulalongkorn University, Thailand
Ladjel Bellatreche	LIAS/ISAE-ENSMA, Université de Poitiers, France
Laurent Anne	Université de Montpellier, LIRMM, CNRS, France
Leandro Antonelli	Universidad Nacional de La Plata, Argentina
Leandro Silva	Universidade Federal de Alagoas, Brazil
Leila Ben Ayed	National School of Computer Science, Tunisia
Leopoldo José Ríos	Universidad Nacional del Nordeste, Argentina
Leticia Flores Pulido	Universidad Autónoma de Tlaxcala, Mexico
Luca Cernuzzi	Catholic University Nuestra Señora de la Asunción, Paraguay
Lydia Boudjeloud-Assala	Université de Lorraine, CNRS, LORIA, France
Luigi Di Caro	University of Turin, Italy

Luis Chiruzzo	Universidad de la República, Uruguay
Luis Corpas Pastor	Universidad de Málaga, Spain
Luis Vives	Universidad Peruana de Ciencias Aplicadas, Peru
Luisa Mich	University of Trento, Italy
Maciej Ogrodniczuk	Institute of Computer Science, Polish Academy of Sciences, Poland
Maguelonne Teisseire	INRAE, France
Manuel Mager	Johannes Gutenberg University Mainz, Germany
Marco Antonio Sobrevilla Cabezudo	Aveni, Pontificia Universidad Católica del Perú, Peru
Marcos Domingues	State University of Maringa, Brazil
Marcos Martinez Galindo	IBM Research, USA
Marcos Quiles	Universidade Federal de São Paulo, Brazil
Maria Clara Gómez Álvarez	Universidad Nacional de Colombia, Colombia
Maresa Schröder	LMU Munich, Germany
Mark Barrueta	Universidad Peruana de Ciencias Aplicadas, Peru
Martin Dominguez	Universidad Nacional de Córdoba, Argentina
Martin Gaedke	Chemnitz University of Technology, Germany
Mathieu Roche	CIRAD, France
Matilde Pato	Instituto Superior de Engenharia de Lisboa, Portugal
Matteo Francia	University of Bologna, Italy
Mattia Prosperi	University of Florida, USA
Mete Celik	Erciyes University, Turkey.
Miguel Alonso	Universidad de La Coruña, Spain
Miguel Nunez-del-Prado	Universidad Peruana de Ciencias Aplicadas, Peru
Miguel Angel Galarreta Valverde	Universidad Nacional de Ingeniería, Peru
Mohamed Jmaiel	University of Sfax, Tunisia
Mourad Oussalah	Nantes University, France
Nandana Mihindukulasooriya	IBM Research AI, USA
Nathalie Abadie	IGN, LaSTIG, France
Nazha Selmaoui-Folcher	ISEA / Université de la Nouvelle Calédonie, France
Nestor Duque Méndez	Universidad Nacional de Colombia, Colombia
Niall Curry	Manchester Metropolitan University, UK
Nicolas Travers	De Vinci Higher Education, De Vinci Research Center, France
Nils Murrugarra-Llerena	University of Pittsburgh, USA
Omar Boussaid	University Lumière Lyon 2, France
Omar Cortes	Instituto Federal do Maranhão, Brazil
Ona de Gibert	University of Helsinki, Finland
Orit Hazzan	Technion - Israel Institute of Technology, Israel

Pablo Fillottrani	Universidad Nacional del Sur, Argentina
Paolo Soda	Università Campus Bio-Medico di Roma, Italy
Parameswari Krishnamurthy	International Institute of Information Technology-Hyderabad, India
Pascal Poncelet, Lirmm	University of Montpellier, France
Paula Brito	Universidade do Porto & LIAAD-INESC TEC, Portugal
Paula Estrella	Universidad Nacional de Córdoba, Argentina
Pawan Kumar	INRIA, France
Peihao Xiang	Florida International University, USA
Pedro Shiguihara	Universidad San Ignacio de Loyola, Peru
Philippe Cudre-Mauroux	University of Fribourg, Switzerland
Philippe Roose	University of Pau and the Adour Region, France
Pierre Zweigenbaum	CNRS, France
Pierrick Bruneau	Luxembourg Institute of Science and Technology, Luxembourg
Qingyu Chen	Yale University, USA
Rafael Santos	Instituto Nacional de Pesquisas Espaciais, Brazil
Rafael Giusti	Universidade Federal do Amazonas, Brazil
Rafael Guzman-Cabrera	Universidad de Guanajuato, Mexico
Ralph Hoch	TU Wien, Austria
Raúl Aguilar Vera	Universidad Autónoma de Yucatán, Mexico
Raúl Vázquez	University of Helsinki, Finland
Ravi Kumar	Google, USA
Rama Chellappa	Johns Hopkins University, USA
Ramón Iván Barraza-Castillo	Universidad Autónoma de Ciudad Juárez, Mexico
René Zuñiga Muñoz	Universidad del Cauca, Colombia
Rezarta Islamaj	National Library of Medicine, USA
Ricardo Landa	Cinvestav, Mexico
Richard Chbeir	Université de Pau et des Pays de l'Adour, France
Robert Leaman	National Institutes of Health, NCBI/NLM/NIH, USA
Rodolfo Zevallos	Pompeu Fabra University, Spain
Ronan Tournier	IRIT, Université Toulouse Capitole, France
Ronald Wilson	University of Florida, USA
Rosa González Ramírez	Universidad de los Andes Chile, Chile
Rosario Medina-Rodríguez	INRIA, France
Ruiwen He	De Vinci Higher Education, De Vinci Research Center, France
Sabrine Mallek	ICN Business School, France
Sadok Ben Yahia	Southern Denmark University, Denmark
Sanjay Kumar Madria	Missouri University of Science & Technology, USA

Sara Renjit	Indian Institute of Information Technology Kottayam, India
Shengtian Sang	Stanford University, USA
Shqiponja Ahmetaj	TU Wien, Austria
Silvia Lizeth Tapia Tarifa	University of Oslo, Norway
Smith Washington Arauco Canchumuni	Pontifical Catholic University of Rio de Janeiro, Brazil
Somnath Banerjee	Tartu Ülikool, Estonia
Sonia Djebali	De Vinci Higher Education, De Vinci Research Center, France
Sophia Ananiadou	University of Manchester, UK
Soto Montalvo	Universidad Rey Juan Carlos, Spain
Stefano Berretti	University of Florence, Italy
Stephan Reiff-Marganiec	University of Derby, UK
T. K. Prasad	Wright State University, USA
Tara Nourivandi	University of South Florida, USA
Tianda Sun	University of York, UK
Tullio Vardanega	University of Padua, Italy
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Valery Solovyev	Kazan Federal University, Russia
Victor Andrés Ayma Quirita	Pontificia Universidad Católica del Perú, Peru
Victor Lomas-Barrie	IIMAS-UNAM, Mexico
Wenyu Song	Harvard Medical School, USA
Wissam Mallouli	Montimage, France
Yi Guo	University of Florida, USA
Yiru Zhang	De Vinci Higher Education, De Vinci Research Center, France
Yousra Hlaoui	University of Tunis El Manar, Tunisia
Yuan An	Drexel University, USA
Yuan Zhao	National Institute of Mental Health, USA
Zhaoyang Wang	Johns Hopkins University, USA
Zhiqiang Wang	De Vinci Higher Education, De Vinci Research Center, France
Zlatko Stapic	University of Zagreb, Croatia
Zoltan Miklos	IRISA, University of Rennes, France
Zoubida Kedad	Universite de Versailles-St-Quentin, France

DISE Program Committee

Carlos Gavidia Calderon	Alan Turing Institute, UK
Carlos Efraín Iñiguez Jarrín	Escuela Politécnica Nacional, Ecuador

Denisse Muñante Arzapalo	ENSIIE & Samovar, France
Essa Imhmed	Eastern New Mexico University, USA
Gandhi Samuel Hernandez Chan	CentroGeo - CONACYT, Mexico
Gonzalo Peraza Mues	Tecnológico de Monterrey, Mexico
Jose Ignacio Panach Navarrete	Universitat de València, Spain
Jose Fabián Reyes-Román	Universitat Politècnica de València, Spain
Otto Parra	Universidad de Cuenca, Ecuador
Sarbagya Shakya	Eastern New Mexico University, USA
Silvia Rueda Pascual	Universitat de València, Spain

Organizing Institutions and Sponsors

Organizing Institutions

Universidad Peruana de Ciencias Aplicadas Peru[3]

Collaborating Institutions

Universidad Nacional Mayor de San Marcos, Peru[4]
De Vinci Pôle Higher Education, De Vinci Research Center, France[5]
Aimara research group, Peru[6]

[3] https://www.upc.edu.pe/en/
[4] https://www.unmsm.edu.pe/
[5] https://www.devinci.fr/
[6] https://www.aimaralab.com/

Contents

Set of Guidelines for the Development of a Web-Based Assistant Focused on the Creation of Dubbing Scripts 1
Renzo Manuel Ramos Ramirez, Orlando Arturo Roca Huapaya, and Eduardo Díaz

ASSIST: A Multi-agentic Framework for Human Computer Interaction in Cultural Heritage Settings 13
Samuel Ramos-Varela, Anmol Guragain, Jaime Bellver-Soler, David Aragon Diaz, Long Lin, and Luis Fernando D'Haro

Geometric Goodness-of-Fit Measure for Interval-Valued Data 27
Dylan Benavides, Oldemar Rodríguez, and Maikol Solís

Multi-objective Optimization for Strategic Shelter Placement: Case Study of Lima's Earthquake Preparedness 45
Soledad Espezúa, Alexandra Sanjinez, Amy Checcllo, and Alexia Ríos

Solving the Qubit Transmission Problem with Biologically Inspired Neural Networks and Reinforcement Learning 60
Ariel Fiuri, Martín Domínguez, Sofía Perón Santana, and Omar Osenda

Horoyah: A Knowledge Graph-Based Framework for Election Monitoring in Low and Middle-Income Countries 79
Gayo Diallo

Semantic Similarity Analysis of Theses Using Sentence Embeddings and Knowledge Graphs 93
Ibeth J. del Pilar Cusi, Lizeth C. Mamani, Gabriela Zuñiga-Rojas, Edwin Alvarez-Mamani, and Harley Vera-Olivera

Measuring Healthcare Accessibility Through Clustering Techniques and Multi-layer Network Analysis 109
Dhananjay Balakrishnan, Svea Drekshagen, Vedant Sahu, Mersedeh Tariverdi, and Miguel Nunez-del-Prado

Intelligent Query System for Consumer Protection Regulations in Perú Using Hierarchical Vector Databases 125
Franco Cárdenas-Martínez and Antonio Arroyo-Paz

Faithfulness and Relevance in Stable Normative Domains: A Comparative Analysis of CAG vs. RAG 137
Elmer A. Collanqui, Maribel E. Coaguila, Anibal Flores, and Honorio Apaza

BERT-Based Approach for Quechua Dialect Identification 154
Christian Aranibar Solaligue and Julio Santisteban

Quechua Speech Synthesizers 167
Edwin Huamán Curo and Ferdinand Pineda Ancco

Quechua Speech Datasets in Common Voice: The Case of Puno Quechua 184
Elwin Huaman, Wendi Huaman, Jorge Luis Huaman, and Ninfa Quispe

Detecting Explicit Content in Spotify Song Lyrics Using Natural Language Processing Techniques 194
Aaron C. A. Navarro Mendoza, Dylan J. Huarcaya Farfan, Fernando A. Usurin Arias, Marcelino Maita-Vargas, and Ariana M. Villegas Suarez

Spam Detection in YouTube Comments Using a Multinomial Naive Bayes Classifier 207
Henry Ivan Condori-Alejo, Deyvis Mamani-Lacuta, and Guina Sotomayor Alzamora

Prompt Engineering in Spanish for Improving Sexism Detection in Tweets 222
María Isabel Limaylla-Lunarejo

Beyond Benchmark Accuracy: Evaluation of Ab Initio MiRNA Detection Models with Genomic Screening Considerations 234
Juan Opazo and Edwin Villanueva

Drug Repurposing for Rare Diseases: A Novel Knowledge Graph Explainable Approach to Identify Drug Candidates 246
Martin Drancé, Fleur Mougin, Akka Zemmari, and Gayo Diallo

Computational Prediction of Intracellular Signaling Behavior via Machine Learning 263
Pamela Romero, Juliette Gourdon, Romain Yvinec, Frederic Jean-Alphonse, and Misbah Razzaq

MeSH Concept Relevance and Knowledge Evolution: A Data-Driven Perspective 280
Jenny Copara, Nona Naderi, Gilles Falquet, and Douglas Teodoro

Fine-Tuning of Small Language Models for Ecological Monitoring: A Comparative Analysis Based on Perplexity 300
Jhon Edilberto Monroy Barrios, Enrique Miguel Paco Cusi, Miguel Angel Portella Huincho, Jeymi Melanie Valdivia Eguiluz, Basurco Zapata Gerald Francis, and Fiorella Jara Rosales

A Variance-Invariance Approach to Better Word Embeddings 314
Zainab Albujasim and Diana Inkpen

Multimodal Sign Language Model 327
Christian Velasquez Borasino, Giorgio Mancusi Barreda, and Rody Vilchez Marin

Implementation of Intelligent User Interfaces and Their Impact on Software Quality: A Systematic Mapping Study 342
Pedro Aguilar-Encarnacion, Carlos Iñiguez-Jarrín, and Julio Sandobalín

Semantic Search for Financial Products Using Embeddings and Pgvector 358
Esteven Calcina-Puma and Antonio Arroyo-Paz

Automatic Debate Generation for Spanish Educational Podcasts Using LLMs 374
Ricardo Lazo Vasquez and José Ochoa-Luna

When One Size Doesn't Fit All: A Study on Intrinsic Iterative Self-correction in LLMs for Hallucination Control 387
Svetlana Esina and Diana Inkpen

Clustering Spatiotemporal Patterns of Infrastructure Investment in Ecuador: A Cantonal-Level Analysis (2008–2022) 403
Helen Magdalena Gómez Medina, Myriam Urbina, María Elena Jerez Calero, Víctor Cuenca, Arcenio Córdova, and Erick Cuenca

Where Is the Research Heading? A Text Mining Analysis of International Business Publications in Peru 419
Hugo Giraldo, Eva Farromeque, Adriana Ticse, Grycel Ticse, Julio Ricardo Moscoso Cuaresma, Miguel Nunez-del-Prado, Juan Eduardo Acosta Mantaro, and Carlos Alberto Azabache Morán

Clustering Analysis of Student Satisfaction–Performance Relationships in Higher Education Using Large-Scale Educational Data 433
Michael Newton Cutipa-Santi, Daisy Crishyely Quelca-Velasquez, Yeni Cutipa-Santi, Renzo Apaza-Cutipa, and Vladimiro Ibañez-Quispe

Predictive Airfare Analysis Based on Machine Learning Models 450
Yasumy Maricely Jallo-Paccaya, Jadyra Ch'aska Choque-Quispe, Thalia Quispe-Mamani, Rebeca Araceli Ccansaya-Soncco, Edwin Alvarez-Mamani, and Harley Vera-Olivera

Mobile App for Detecting Work-Related Stress Using Acoustic Biomarkers and Machine Learning 468
Claudia Lucia Rozas Gamero, Oscar Heredia Miranda, Eduardo Díaz, and Edgar Ceh-Varela

Evaluation of Machine Learning Algorithms to Estimate the Prevalence of Anemia in Children at the District Level in Peru 480
Anthony Mamani, Daniel A. Gutierrez-Pachas, and Luis Rosero

Socioeconomic Factors Associated with Depression in Peru: A Machine Learning-Based Study Using ENDES 2016–2023 495
Fernando Bendezú-Galarza and Miguel Nunez-del-Prado

Conjecture Extraction for Proof Autoformalization 510
Simon Sorg, Wenda Li, and Soumya Banerjee

Counting Avocados in Orchard Videos Using a YOLO and BoT-SORT-Based Architecture 522
Diego Albitres Blondet, Franco Galindo Alvarez, and Carlos Fernando Montoya Cubas

Guidelines for Designing Mobile Applications for Comprehensive Support of Stomach Cancer Patients in Lima: Medical and Spiritual Assistance 534
Martin Cifuentes Mellado, Jean Morin Fuentes, Eduardo Díaz, and Carlos Iñiguez-Jarrín

Author Index 551

Set of Guidelines for the Development of a Web-Based Assistant Focused on the Creation of Dubbing Scripts

Renzo Manuel Ramos Ramirez, Orlando Arturo Roca Huapaya, and Eduardo Díaz(✉)

Universidad Peruana de Ciencias Aplicadas, Prolongación Primavera 2390, Lima 15023, Peru
{U202113745,U201919742}@upc.edu.pe

Abstract. Dubbing studios in Latin America face challenges in creating translated scripts due to linguistic inconsistencies, limited cultural adaptation, and extensive manual processes. This article proposes ten guidelines for designing a web application for script preparation in dubbing projects. The guidelines are based on 31 scientific articles and are organized into two categories: (i) Guidelines for interaction design and user experience, and (ii) Guidelines for the architecture and technologies of the web application. To provide further clarity, an illustrative example of the proposed guidelines is presented. Additionally, an experiment was conducted with 20 participants to measure satisfaction with the web application designed according to the proposed guidelines, yielding positive results. This research may be of interest to software designers, dubbing studios, and audiovisual localization researchers seeking to improve productivity and narrative quality in dubbing workflows.

Keywords: Web application · generative artificial intelligence · translated scripts · automatic transcription · machine translation · web assistant

1 Introduction

Dubbing script preparation faces recurrent bottlenecks in linguistic adaptation, cultural adjustments, lip-sync, and narrative coherence, which extend timelines even when translation quality is maintained. Illustrative cases include Avatar (three months to resolve jargon and colloquialisms) [1] and Maya the Bee: The Honey Games (better synchronization and cohesion in Peruvian Spanish than in European Spanish, with longer production) [2]. A further factor is the lack of uniform standards in some markets, for example the English-speaking one, which complicates lip-sync, prosody, and fluency [48]. Current AI tools are insufficient for these workflows: Heygen speeds multilingual video with lip-sync but limits video length and lacks cultural adaptation [3], while Vink uses Whisper for diarization and transcription but still struggles with synchronization and adaptation [4].

This work proposes ten guidelines derived from 31 studies. Interaction/UX: text-based conversational interfaces, color/contrast accessibility, mixed filtering and search,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 1–12, 2026.
https://doi.org/10.1007/978-3-032-20322-9_1

rich-text editing, adjustable-speed video playback. Architecture/technology: LLM integration, speech-to-text, diarization, Transformer-based MT, AI-assisted translation. Unlike general HCI guidance for editing or subtitling, the proposal targets dubbing needs such as character-based timecodes, lip-sync, and culturally appropriate adaptation for Latin America, making it applicable to studios in emerging markets. The paper includes an illustrative implementation and is organized as follows: Sect. 2 (state of the art), Sect. 3 (guidelines), Sect. 4 (example), Sect. 5 (early validation), and Sect. 6 (conclusions and future work).

2 State of Art

This section reviews prior work on the technologies needed to build a web assistant that transcribes and generates translations for dubbing scripts. We conducted a Targeted Literature Review (TLR), non-systematic and informative, to select the most relevant studies and reduce selection bias. Searches were run in Scopus using the query: ("speaker diarization" OR "speech segmentation" OR "speaker identification") AND ("language model" OR "large language model" OR "LLM" OR "specialized domain"). Sixty articles were initially retrieved. After applying inclusion criteria (use of speaker diarization for voice separation; application of LLMs in specialized domains) and exclusion criteria (no focus on accuracy, efficiency, or applicability; unrelated to transcription, dubbing, or specialized text generation), ten articles remained for detailed analysis. The final set was classified into two groups:

2.1 Speaker Diarization Technologies

Speaker diarization. It segments an audio stream by speaker to answer who spoke when without prior identities, and it supports transcription, turn-taking analysis, and discourse studies [46]. Recent work shows consistent gains across settings: Pan et al. introduce ATC-SD Net for noisy aeronautical radiotelephony, reporting 6.3% DER on ATCSPEECH and outperforming VBx [5]; Khoma et al. present a PyAnnote-based supervised system where clustering beats segment classification (15.52% vs. 20.14% DER) [6]; Lyu et al. build a multilingual real-time pipeline on Whisper with WDER of 2.68% for two speakers and 6.96% for three on political interviews [7]; Aung et al. enhance EEND with a multiscale decoder, reaching 4.37% DER and surpassing i-/x-vector baselines [8]; Xylogiannis et al. fuse acoustic and spatial cues, reducing DER by 2–3% on AVLAB and SpeaD-M3C [9]. Overall, the literature emphasizes neural architectures [5, 8], multilingual robustness [7], and multisensor fusion [9]. Our project adopts these insights to design a computationally efficient diarization module suitable for web-based workflows and varied conversational contexts.

2.2 Large Language Models

Large Language Models (LLMs) are Transformer-based neural networks trained for next-token prediction and applied to generation, translation, and summarization, often outperforming traditional approaches [47]. Evidence from machine translation research

shows that LLMs can approach offline quality in simultaneous settings through adaptive read–write training and supervised fine-tuning data design [10]. Idiom handling improves when aligning figurative meaning via Semantic Idiom Alignment and LLM-based Idiom Alignment, with human evaluations favoring semantic alignment [11]. Chinese idiom translation remains challenging, and even top systems still make substantial errors while common metrics correlate poorly with human judgments [12]. For low-resource Indian languages, evaluations using sentiment and semantic analyses compare GPT-4o, Gemini, and Google Translate against expert references to assess quality [13]. Iterative Bilingual Understanding further refines translations by generating and improving bilingual context, yielding measurable COMET gains across domains [14]. Taken together, these results indicate strong applicability of LLMs to translation, including real-time and idiomatic use cases, while highlighting the need for robust evaluation and control strategies [10–14].

3 Definition of the Proposed Guidelines

This section presents a set of 10 guidelines derived from the literature review for designing a web application for the preparation of dubbing scripts. The guidelines are based on the analysis of 31 scientific articles. Each guideline is supported by scientific studies that describe the necessary considerations for designing a web application for dubbing script preparation. The analyzed studies address aspects such as interface design, the use of services like diarization, artificial intelligence, Transformers, and Large Language Models (LLMs). These guidelines are grouped into two main categories: (i) Guidelines for interaction design and user experience, and (ii) Guidelines for the architecture and technologies of the web application. Each guideline is identified with the prefix "G" followed by a number (for example, G1, G2, G3, etc.).

3.1 Guidelines for Interaction Design and User Experience

G1: Conversational Interfaces. Studies on text-based conversational interfaces for web assistants indicate measurable UX gains on complex tasks when designs provide clear structure, visual feedback, and accessible controls. Kuang et al. compare voice and text assistants in UX evaluations and report that text systems are perceived as more efficient, easier to use, and less intrusive in analytical contexts, supported by visual clarity and the option to review asynchronously [31]. Silva and Canedo, in a systematic review, identify dialog-style message structuring, visible interaction history, and upfront transparency about limitations as core factors [32]. Srinivas et al. examine proactive LLM-based assistants for assessment tasks in complex cognitive processes and find that proactivity, user control over interaction flow, and adjustable assistant behavior improve perceived usefulness, trust, and control [33]. In conclusion, use a persistent, text-based interface embedded in the workflow with clear visual feedback and simple turn-taking. Provide a visible, revisitable history, adjustable proactivity, and upfront transparency about limits. This enables reviewing, repeating, or modifying steps improving efficiency, trust, and perceived control.

G2: Colors and Text. Research on color schemes and visual accessibility in subtitling and dubbing contexts indicates that careful theme and contrast choices improve

readability, reduce eye strain, and support inclusive, long-duration work. Fan et al. report that under negative polarity (dark mode) yellow on black yields the lowest visual fatigue, while red and some green hues increase discomfort (eye-tracking, blink rate) [34]. Szarkowska and Boczkowska find that color-coding subtitles by language does not harm comprehension and can increase immersion for certain viewers [35]. Cheng, Vahdat, and Lin introduce the Primary Color Difference metric and show that high-contrast text–background pairs, such as green on black or white, improve recognition efficiency and reduce cognitive load on LCDs, especially at higher contrast levels [36]. In conclusion, it is recommended that the web assistant for dubbing script creation implement a dark mode with text in high-contrast colors such as yellow, green, or white, avoiding colors like red that increase visual fatigue.

G3: Dynamic Filters and Classification Systems. Evidence shows that navigation improves when interfaces combine free-text search, hierarchical categories, and tags with clear interactions, immediate feedback, and flexible user control. Niu et al. report that faceted filters should expose relevant attributes, keep active selections visible at all times, and prevent combinations that yield empty results, users prefer filters that can be applied and removed intuitively without interrupting the workflow [37]. Melenhorst et al. find that user-generated tags support flexible, personalized classification when they are visible in the UI, allow multiple combinations, and align with natural language, which suits scene, emotion, or pending-edit tagging in script work [38]. Mahdi et al. review dynamic filter design and support real-time application, per-option result previews, and the pairing of hierarchies with free-form tags to accommodate different profiles and large editable datasets [39]. In conclusion, it is recommended that the web application for dubbing script creation implement a mixed filtering system composed of textual search, hierarchical categories, and customizable tags, allowing filters to be applied dynamically, displaying active filters at all times, and enabling easy removal.

G4: Text Editing. Studies on rich-text editors in collaborative settings show efficiency gains in writing and review when features such as anchored comments, version history, and change tracking are implemented with visual clarity, authorship control, and accessibility. Shulgina et al. report that comments attached to specific fragments significantly improve writing performance, whereas using tracked changes in isolation is less effective [40]. Das et al. design accessible collaboration techniques for visually impaired users, adding auditory representations of comments and suggested edits; in tests with 48 screen-reader users these functions improved understanding of changes, increased perceived participation, and reduced barriers [41]. Birnholtz and Ibara analyze the social dynamics of collaboration, finding that visible edits influence interpersonal relations, many contributors prefer to justify modifications through comments, and exposing authorship with controlled accept or reject actions supports coordination [42]. In conclusion, it is recommended that the web application for dubbing script creation implement a rich-text editor that allows the user to apply standard formatting (bold, italics, underline), insert comments anchored to specific script fragments, conduct reviews with visible change tracking, and access a version history with author identification.

G5: Video Player. Studies on web video players show efficiency gains in editing and dubbing when interfaces include speed control, subtitles, and tight script synchronization, implemented with intuitive and accessible controls. Christel et al. develop a player

with synchronized transcription that lets users navigate via line-linked text; tests show faster retrieval of relevant segments and better comprehension, with direct jumps from script to scene reducing cognitive load [43]. Valor Miró et al. find that simple, functional subtitle editors help users adjust auto-generated captions without technical overhead, cutting editing time by up to 70% compared to manual transcription [44]. Pantula compares accessible web players under WAI guidelines and reports that recent systems often integrate speed control, editable subtitles, and assistive-tech synchronization, improving experiences for users with visual or hearing impairments [45]. In conclusion, it is recommended that the web application for dubbing script creation implement a video player with adjustable playback speed, synchronized subtitle display, direct editing capabilities for the text, and script-based navigation. These functions should be accessible, intuitive, and provide clear feedback, allowing users to quickly move between scenes, review content with precision, and maintain an efficient dubbing-focused workflow.

3.2 Guidelines for the Architecture and Technologies of the Web Application

G6: LLM Integration. Studies on controlled, reliable LLM use in specialized applications point to three consistent practices. Woollaston et al. present TAMMY, an LLM-powered chatbot for EFL translation tasks, showing effectiveness when interaction flows are constrained, and the model operates at low temperature to limit variability [15]. Raunak et al. propose GPT-4 for automatic post-editing, demonstrating significant improvements over baseline MT while stressing the need for human oversight and careful prompt design to avoid hallucinations or unnecessary edits [16]. Lee et al. introduce a Modular Prompted Chatbot (MPC) that maintains long-term consistency by decomposing services into modules and applying techniques such as chain-of-thought and specialized prompting, removing the need for costly fine-tuning in many cases [17]. In conclusion, integrate LLMs through a modular, asynchronous architecture with the model as a decoupled service (API or queue) in a staged pipeline that includes reasoning, contextual prompting, and post-editing to enrich inputs and maintain coherence without fine-tuning. Use a middleware flow manager to orchestrate calls, adjust creativity and variability, and route domain-specific prompts.

G7: Automatic Speech-to-Text Transcription. Prior work supports a two-stage pipeline for spoken content: first transcribe, then translate. Zhang formalizes this separation, noting that structured text simplifies semantic, grammatical, and syntactic handling compared with operating directly on audio [18]. Siddique Latif et al. show that Transformer-based ASR outperforms RNN and CNN approaches, better managing variability, accents, and complex linguistic patterns required for high-quality transcripts [19]. Radford et al. present Whisper, a multilingual model trained on 680,000 h that transcribes multiple languages without extra training and outperforms alternatives such as wav2vec 2.0 across diverse recordings, with near-human results in tests [20]. This evidence motivates adopting ASR as an essential first step in multilingual workflows, followed by MT on the transcribed text. In conclusion, automatic Speech-to-Text transcription is recommended as an essential step for capturing dialogues in multilingual contexts. Furthermore, it is suggested that the web assistant for dubbing script preparation implement a two-stage workflow: (i) transcription of audio using advanced models such as Whisper, and (ii) translation of the transcribed text.

G8: Integration of Diarization Services. Studies on integrating external diarization into web transcription systems point to joint or modular pipelines that handle multiple speakers and languages. Cheng et al. present a simultaneous ASR–diarization approach using Conformer, with information sharing across stages and Seq2Seq-TSVAD for parallel target-speaker activity detection, avoiding multiple passes [21]. Vachhani et al. use Whisper for transcription and language ID, adding ECAPA-TDNN accent detection, multi-window analysis, and DOVER-Lap fusion to improve multilingual accuracy [22]. Pappala et al. combine fine-tuned Whisper for Hindi/Marathi with Pyannote.audio for speaker segmentation in a reproducible, service-oriented pipeline that also accommodates sentiment analysis [23]. Collectively, these works support modular, decoupled integration of transcription, language detection, and diarization services exposed via APIs to scale across complex, multilingual audio. In conclusion, it is recommended to integrate automatic speaker diarization services in a modular and decoupled manner, employing existing tools such as Whisper (for transcription and language detection) and Pyannote.audio (for speaker segmentation). These services should be invoked through REST APIs or other asynchronous mechanisms such as message queues, facilitating their integration into a scalable architecture. This strategy ensures that the system can produce scripts correctly segmented by speaker, even in complex contexts with overlapping voices or language alternation.

G9: Transformer-Based Machine Translation. Reviews and evaluations highlight the centrality of Transformer models in MT and their advantages over statistical and RNN approaches, with better semantic context capture and fluency [24]. For dubbing, these properties help preserve rhythm, intent, and narrative cohesion in dialogue. Comparative studies show GPT-4 achieves accuracy comparable to junior human translators, though it can miss cultural and expressive nuances that still require human refinement [25]. Additional analyses report GPT-4 is competitive with commercial MT systems, especially when prompting is carefully designed [26]. In conclusion, it is recommended that the web assistant for dubbing script preparation implement a transformer-based machine translation system, considering the use of GPT-4 to generate high-quality initial translations and applying assisted post-editing to ensure semantic fidelity.

G10: Human-in-the-Loop Translation. Studies on AI-assisted MT with human intervention for professional, confidentiality-sensitive settings converge on hybrid workflows. Yuksel et al. combine LLMs with post-editing: the model proposes multiple hypotheses, scores them with COMETQE and GEMBA, and routes only critical segments to human editors; accepted fixes are reused via active learning [27]. Chatzitheodorou et al. design a secure pipeline with pseudo-anonymization before MT, human review on anonymized text, and reinsertion of original entities after validation, which is pertinent for unpublished scripts [28]. Han introduces HilMeMe, a human-centric MT metric focused on idioms and multi-word expressions, classifying errors by type and severity to guide fluent and adequate revisions [29]. Yang et al. present an LLM pipeline that generates a draft, then refines it with examples and feedback using in-context learning; past corrections are stored for reuse, and the system can select between draft and revision through model self-evaluation [30]. Collectively, these works support modular MT pipelines where the model drafts, humans correct targeted segments, and the system learns from edits

while preserving data security. In conclusion, it is recommended that the web application implement a Human-in-the-Loop-based machine translation workflow, where the AI generates a first draft of the script and this is refined through human intervention. To achieve this, it is suggested that the components responsible for translation, review, correction storage, and improvement suggestions be developed as decoupled microservices, invoked via REST APIs.

Table 1 summarizes the ten recommendations proposed for designing a web application for creating dubbing scripts.

Table 1. Summary of Recommendations

Category	Guidelines	Details
Guidelines for Interaction Design and User Experience	G1	Conversational interfaces
	G2	Colors and text
	G3	Dynamic filters and classification systems
	G4	Text editing
	G5	Video player
Guidelines for the Architecture and Technologies of the Web Application	G6	LLM integration (e.g., GPT-4)
	G7	Automatic speech-to-text transcription
	G8	Integration of diarization services
	G9	Transformer-based machine translation
	G10	Translation with human-in-the-loop

4 Illustrative Example of the Proposed Guidelines

This section presents an illustrative example of seven guidelines for designing a web application for dubbing script preparation, demonstrating how these guidelines enhance the process of creating dubbing scripts.

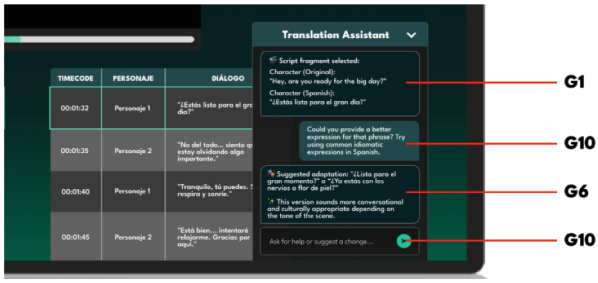


Fig. 1. Assistant Conversational Interface with AI-Assisted Cultural Adaptation.

Figure 1 shows a chatbot responding to a request to translate a script fragment into Spanish. It illustrates G6 through the use of an LLM to handle user queries, G10 by involving the user in reviewing and refining the draft rather than accepting an automatic result, and G1 in the UI/UX through a text-based interface with preserved message history.

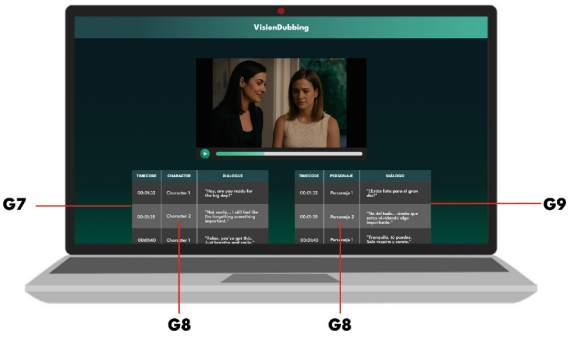


Fig. 2. Interface Showing the Results of Transcription, Translation, and Diarization of Dubbing Scripts Produced by the Web Assistant.

Figure 2 presents two tables: the transcription (G7) and the translated script (G9). Both list "Timecode," "Character," and "Dialogue." The separation of entries is enabled by diarization (G8), which segments the audio, identifies speakers, and provides the corresponding durations.

5 Early Validation

This section describes an early validation performed to verify the applicability of the proposed guidelines for the development of a dubbing web assistant. This prototype was created following key recommendations such as G1 (Conversational interfaces), G2 (Colors and text), G6 (Integration of LLMs), G7 (Automatic speech-to-text transcription), G8 (Integration of diarization services), G9 (Machine translation based on transformers), and G10 (Translation with human-in-the-loop). This first version implemented main functionalities like transcription, speaker diarization and translation. The validation involved twenty participants with professional experience in dubbing workflows, ranging from junior practitioners to professionals with moderate industry expertise, who performed predefined tasks using the web assistant. Each subject interacted with the prototype by uploading an English-language or French-language MP4 file and completing the process to obtain a translated and segmented script in Spanish. The system allowed participants to visualize and revise speaker-specific segments, edit translations, and export the script. When subjects finished solving the experimental cases, they completed a satisfaction survey. The satisfaction was measured online using a 5-point Likert scale questionnaire based on the framework developed by Moody's [49], which defined a framework (based on the work of Lindland et al. [50]) to evaluate satisfaction in terms of Perceived Ease

to Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU). This framework has been previously validated and is widely used. The possible answers for each statement in the PEOU, PU, and ITU questionnaire are: Totally disagree, Fairly disagree, Neutral, Fairly agree, and Totally agree. A numerical value is provided to each statement from 1 (Totally disagree) to 5 (Totally agree). Six questions were defined to measure PEOU; the metric was calculated by adding the numerical values of the answers and classifying the result into a rank of five possible values: Rank 1–6: Totally disagree, Rank 7–12: Fairly disagree, Rank 13–18: Neutral, Rank 19–24: Fairly agree, Rank 25–30: Totally agree. For example, if a subject answers 5 questions with Totally agree and 1 question with Neutral in PU, the result of this metric will be 28 (Totally agree). Six questions were defined to measure PU; the metric was calculated by adding the numerical values of the answers that each subject filled in through the six questions, and the result of this addition is classified into a rank with the five possible options: Rank 1–8: Totally disagree, Rank 9–16: Fairly disagree, Rank 17–24: Neutral, Rank 25–32: Fairly agree, Rank 33–40: Totally agree. Six questions were defined to measure ITU; the metric was calculated by adding the numerical values of the answers and classifying the result into a rank of five possible values: Rank 1–2: Totally disagree, Rank 3–4: Fairly disagree, Rank 5–6: Neutral, Rank 7–8: Fairly agree, Rank 9–10: Totally agree. Fig. 3 shows the satisfaction results.

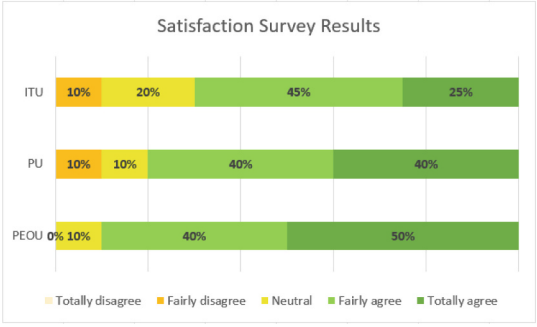


Fig. 3 Results of satisfaction

Figure 3 shows the distribution of subjects responses across of PEOU, PU, and ITU. A large majority of respondents selected "Fairly agree" or "Totally agree" in each category Specifically, 90% of subjects agreed that the system was easy to use (PEOU), 80% considered it useful for dubbing workflows (PU), and 70% expressed their intention to use it in future tasks (ITU). These results demonstrate encouraging early adoption potential and support the relevance of the proposed design guidelines.

6 Conclusions and Future Work

This work proposes ten guidelines for designing a web application for dubbing script preparation, derived from 31 studies. G1 covers conversational interfaces; G2 colors and text; G3 dynamic filters and classification; G4 rich-text editing; G5 the video

player; G6 LLM integration; G7 speech-to-text transcription; G8 speaker diarization; G9 Transformer-based machine translation; and G10 AI-assisted translation with human oversight. We present an illustrative implementation that integrates G1, G2, G6, G7, G8, G9, and G10, showing how these guidelines improve the workflow. An early study with 20 participants reports positive satisfaction results. Future work includes adding guidelines grounded in new evidence and running broader empirical evaluations of effort and satisfaction in dubbing scenarios.

References

1. Cutando Plumed, A.: Traducción de guiones y tiempos de producción en la industria audiovisual: un análisis del caso de la saga Avatar (2023). https://hdl.handle.net/10234/203484
2. Guevara, L. & Quiroz, M. (2022). Comparación de versiones dobladas en distintos mercados: Un análisis de "Maya the Bee: The Honey Games" en español peruano y español peninsular. https://hdl.handle.net/20.500.12692/129174
3. Miquel-Vergés, J.: Heygen: Soluciones de inteligencia artificial para la creación rápida de vídeos multilingües con sincronización labial (2024). https://doi.org/10.31637/epsir-2024-358
4. Tolle, H., et al.: From voice to ink (Vink): development and assessment of an automated, free-of-charge transcription tool. BMC. Res. Notes. **17**(1), 95 (2024). https://doi.org/10.1186/s13104-024-06749-0
5. Pan, W., Wang, Y., Zhang, Y., Han, B.: ATC-SD net: radiotelephone communications speaker Diarization network. Aerospace. **11**(7), 599 (2024). https://doi.org/10.3390/aerospace11070599
6. Khoma, V., Khoma, Y., Brydinskyi, V., Konovalov, A.: Development of supervised speaker diarization system based on the PyAnnote audio processing library. Sensors. **23**, 2082 (2023) https://www.mdpi.com/1424-8220/23/4/2082
7. Lyu, K.M., Lyu, R., Chang, H.T.: Real-time multilingual speech recognition and speaker diarization system based on whisper segmentation. PeerJ Comput. Sci. **10**, e1973 (2024). https://doi.org/10.7717/peerj-cs.1973
8. Aung, M.A.A., Pa, W.P., Naing, H.M.S.: End-to-end neural Diarization for unknown number of speakers with multi scale decoder. Int. J. Intell. Eng. Syst. **17**(5), 870–881 (2024). https://doi.org/10.22266/ijies2024.1031.66
9. Xylogiannis, P., Vryzas, N., Vrysis, L., Dimoulas, C.: Multisensory fusion for unsupervised spatiotemporal speaker Diarization. Sensors. **24**(13), 4229 (2024). https://doi.org/10.3390/s24134229
10. Fu, B., et al. (2025). LLMs Can Achieve High-quality Simultaneous Machine Translation as Efficiently as Offline. https://doi.org/10.18653/v1/2025.findings-acl.1045
11. Donthi, S., et al.: Improving LLM Abilities in Idiomatic Translation (2025). https://aclanthology.org/2025.loreslm-1.13/
12. Yang, C., Dou, Y., Heineman, D., Wu, X., & Xu, W.: Evaluating LLMs on Chinese Idiom Translation(2025). https://doi.org/10.48550/arXiv.2508.10421
13. Chandra, R., Chaudhari, A., & Rayavarapu, Y. (2025). An Evaluation of LLMs and Google Translate for Translation of Selected Indian Languages via Sentiment and Semantic Analyses. https://doi.org/10.1109/ACCESS.2025.3585629
14. Chen, A., et al.: LLM-based Translation Inference with Iterative Bilingual Understanding (2025). https://doi.org/10.18653/v1/2025.findings-acl.867

15. Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., Ogata, H.: TAMMY: supporting EFL translation practice with an LLM-powered Chatbot. In: International Conference on Computers in Education (2024). https://doi.org/10.58459/icce.2024.4901
16. Raunak, V., Sharaf, A., Wang, Y., Awadallah, H. H., & Menezes, A.: Leveraging GPT-4 for Automatic Translation Post-Editing (2023). https://doi.org/10.48550/arXiv.2305.14878
17. Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., Lee, K.: Prompted LLMs as Chatbot modules for long open-domain conversation. Find. Assoc. Comput. Linguis.: ACL. **2023**, 4536–4554 (2023). https://doi.org/10.18653/v1/2023.findings-acl.277
18. Zhang, J.: Bridging speech recognition and machine translation: two-stage approaches for spoken language translation. IEEE Access. **10**, 98765–98778 (2022). https://doi.org/10.1109/ACCESS.2022.9876543
19. Latif, S., Rana, R., Qadir, J., Epps, J.: Survey of deep learning techniques for automatic speech recognition. IEEE Access. **7**, 19143–19165 (2020). https://doi.org/10.1109/ACCESS.2020.2968524
20. Radford, A., Jain, S., Mielke, J., Ramesh, A., Kim, J. W., & Sutskever, I.: Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI (2022). https://cdn.openai.com/papers/whisper.pdf
21. Cheng, M., Wang, W., Zhang, Y., Qin, X., & Li, M.: Target-Speaker Voice Activity Detection via Sequence-to-Sequence Prediction (2022). doi:https://doi.org/10.1109/ICASSP49357.2023.10094752
22. Vachhani, B., Singh, D., Lawyer, R.: Multi-resolution approach to identification of spoken languages and to improve overall language Diarization system using whisper model. INTERSPEECH. **2023**, 1993–1997 (2023). https://doi.org/10.21437/Interspeech.2023-1354
23. Papala, G., Ransing, A., Jain, P.: Sentiment analysis and speaker Diarization in Hindi and Marathi using using Finetuned whisper. Scalable Comput. Pract. Experience. **24**(4), 835–846 (2023). https://doi.org/10.12694/scpe.v24i4.2248
24. Tan, Z., et al.: Neural machine translation: a review of methods, resources, and tools. AI Open. **1**, 5–21 (2020). https://doi.org/10.1016/j.aiopen.2020.11.001
25. Yan, J., et al.: Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels (2024). https://doi.org/10.48550/arXiv.2411.13775
26. Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z.: Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine (2023). https://doi.org/10.48550/arXiv.2301.08745
27. Yuksel, K. A., Gunduz, A., Anees, A. B., & Sawaf, H.: Efficient Machine Translation Corpus Generation: Integrating Human-in-the-Loop Post-Editing with Large Language Models (2025). https://doi.org/10.48550/arXiv.2502.12755
28. Chatzitheodorou, K., Escrivá, M.Á.G., Lacal, C.G.: Machine translation of anonymized documents with human-in-the-loop. In: Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pp. 429–436 (2023) https://aclanthology.org/2023.eamt-1.41/
29. Han, L.: HilMeMe: A Human-in-the-Loop Machine Translation Evaluation Metric Looking into Multi-Word Expressions (2022). https://doi.org/10.48550/arXiv.2211.05201
30. Yang, X., Zhan, R., Wong, D. F., Wu, J., & Chao, L. S.: Human-in-the-loop Machine Translation with Large Language Model (2023). https://doi.org/10.48550/arXiv.2310.08908
31. Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., Shinohara, K.: Collaboration with conversational AI assistants for UX evaluation: questions and how to ask them (voice vs. text). In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2023, April). https://doi.org/10.1145/3544548.3581247
32. Silva, G.R.S., Canedo, E.D.: Towards user-centric guidelines for chatbot conversational design. Int. J. Human–Comput. Interact. **40**(2), 98–120 (2024). https://doi.org/10.1080/10447318.2022.2118244

33. Srinivas, V., et al. (2025). Substance over Style: Evaluating Proactive Conversational Coaching Agents. https://doi.org/10.48550/arXiv.2503.19328
34. Fan, Q., Xie, J., Dong, Z., Wang, Y.: The effect of ambient illumination and text color on visual fatigue under negative polarity. Sensors. **24**(11), 3516 (2024). https://doi.org/10.3390/s24113516
35. Szarkowska, A., Boczkowska, J.: Colour coding subtitles in multilingual films – a reception study. Perspectives. **30**(3), 520–536 (2022). https://doi.org/10.1080/0907676X.2020.1853186
36. Cheng, Z., Vahdat, V., & Lin, Y.: A novel approach to study the effect of font and background color combinations on text recognition efficiency on LCDs (2018). arXiv. https://doi.org/10.48550/arXiv.1812.08842
37. Niu, X., Fan, X., Zhang, T.: Understanding faceted search from data science and human factor perspectives. ACM Trans. Inf. Syst. **37**(2), 1–27 (2019). https://doi.org/10.1145/3284101
38. Melenhorst, M., Hoekstra, R., Koolen, M.: Exploring user tagging for supporting search in streaming video services. J. Assoc. Inf. Sci. Technol. **69**(4), 505–517 (2018). https://doi.org/10.1002/asi.23981
39. Mahdi, M.N., Ahmad, A.R., Natiq, H., Subhi, M.A., Qassim, Q.S.: Comprehensive review and future research directions on dynamic faceted search. Appl. Sci. **11**(17), 8113 (2021). https://doi.org/10.3390/app11178113
40. Shulgina, G., Costley, J., Shcheglova, I., Zhang, H.: Online peer editing: the influence of comments, tracked changes and perception of participation on students' writing performance. Smart Learn. Environ. **11**(1), 30 (2024). https://doi.org/10.1186/s40561-024-00315-8
41. Das, M., Piper, A.M., Gergle, D.: Design and evaluation of accessible collaborative writing techniques for people with vision impairments. ACM Trans. Comput.-Hum. Inter. (TOCHI). **29**(2), 1–42 (2022). https://doi.org/10.1145/3480169
42. Birnholtz, J.P., Ibara, S.: Tracking changes in collaborative writing: edits, visibility and group maintenance. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW), pp. 809–818 (2012). https://doi.org/10.1145/2145204.2145325
43. Christel, M.G., Stevens, S.M., Maher, B.S., Richardson, J.: Enhanced exploration of oral history archives through processed video and synchronized text transcripts. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1333–1342 (2010). https://doi.org/10.1145/1873951.1874215
44. Valor Miró, J., Silvestre-Cerdà, J.A., Ferrer, M.A.: Post-editing of automatic subtitles: quality and usability evaluation. Speech Comm. **74**, 61–71 (2015). https://doi.org/10.1016/j.specom.2015.09.006
45. Pantula, M.: A study on media players from an accessibility perspective. Int. J. Comput. Aided Eng. Technol. (IJCAET). **18**(4), 2023 (2023). https://doi.org/10.1504/IJCAET.2023.131920
46. Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S.: A review of speaker diarization: recent advances with deep learning. Comput. Speech Lang. **72**, 101317 (2022). https://doi.org/10.1016/j.csl.2021.101317
47. Zhao, W. X., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2) (2023). https://doi.org/10.48550/arXiv.2303.18223
48. Miggiani, G.S.: English-language dubbing: challenges and quality standards of an emerging localisation trend. J. Specialised Translation. (36), 2–25 (2021). https://doi.org/10.26034/cm.jostrans.2021.054
49. Moody, D. L.: The method evaluation model: a theoretical model for validating information systems design methods (2003). https://aisel.aisnet.org/ecis2003/79
50. Lindland, O.I., Sindre, G., Solvberg, A.: Understanding quality in conceptual modeling. IEEE Softw. **11**(2), 42–49 (1994). https://doi.org/10.1109/52.268955

ASSIST: A Multi-agentic Framework for Human Computer Interaction in Cultural Heritage Settings

Samuel Ramos-Varela(✉), Anmol Guragain, Jaime Bellver-Soler, David Aragon Diaz, Long Lin, and Luis Fernando D'Haro(✉)

Speech Technology and Machine Learning Group, E.T.S.I de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
{s.rvarela,anmol.g,jaime.bellver,long.lin,luisfernando.dharo}@upm.es, daviarag@ucm.es

Abstract. This paper presents ASSIST-AI, a novel multi-modal multi-agentic framework designed to enhance human-computer interaction in cultural heritage environments through advanced AI technologies. Our system integrates computer vision, natural language processing, and speech technologies to create context-aware conversational agents for museum settings. The framework combines Retrieval-Augmented Generation (RAG) with automatic Point of Interest (PoI) detection, personalized user profiling, and real-time multi-modal interaction capabilities. We demonstrate significant improvements in user engagement through adaptive personalization mechanisms that leverage attention schema theory and contextual awareness. Evaluation with 30 participants across major Spanish museums shows enhanced visitor experience and knowledge retention. The system achieves high accuracy in artwork identification, sub-second response times for multi-modal queries, and supports real-time interaction in Spanish and English with adaptive complexity based on user expertise levels.

Keywords: Multi-Modal AI · Multi-Agentic Systems · Cultural Heritage · Human-Computer Interaction · Vision-Language Models · Conversational AI

1 Introduction

Cultural heritage sites such as museums and historical monuments have become fertile testbeds for HCI (Human Compuret Interaction) innovations that blend physical and digital interactions through multimodal interfaces [6]. Multimodal and agent-based interaction frameworks offer advanced information services that adapt to user context and preferences, enabling engaging experiences within these public spaces [1]. Research has explored mobile and wearable technologies, such as multimodal mobile museum guides, to provide inclusive and interactive tours for diverse audiences [20]. Context-aware systems leveraging location-sensing agents have personalized visitor experiences by dynamically deploying

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 13–26, 2026.
https://doi.org/10.1007/978-3-032-20322-9_2

relevant content based on visitor position [8]. The growing field of image-based artwork localization (IAL) and Artwork Point-of-Interest Detection (APoID) demonstrates the feasibility of non-intrusive indoor localization techniques for guiding visitors to exhibits [7]. Additionally, speech interfaces have emerged as a hands-free modality, allowing visitors to navigate virtual or physical museum spaces through spoken commands, including voice-guided wayfinding and group-synchronized tours [15,17,18].

Evolving visitor expectations for personalized, context-sensitive interactions underscore the need to design systems that fulfill individual motivations and enhance engagement [11,22]. Applying self-determination theory to museum interfaces has demonstrated that satisfying visitor needs for autonomy, competence, and relatedness can significantly boost motivation and sustained interaction [14]. Augmented reality guided tours have been shown to mediate between everyday visitor engagement and curated exhibition knowledge, facilitating deeper learning experiences [19]. In addition, context-aware museum guide agents have been developed to address the challenge of providing relevant and timely information, thus aligning technological objectives with visitor goals [9].

In this work, we present a novel HCI framework: VisualRAG + PoI + Speech that integrates visual retrieval-augmented generation for artwork identification, image-based PoI detection and a conversational speech interface for hands-free, natural interaction. In our framework, an agent is a cohesive software component composed of one or more tools and routines that perform a specific task or domain function (e.g. vision, dialogue, moderation). Each agent maintains a local state, exposes a communication interface, and can invoke internal or external tools. Agents can interact, coordinate, and share information, forming a multi-agentic architecture.

By unifying these modalities, our approach creates a seamless hybrid experience in cultural heritage settings, enhancing accessibility, personalization, and engagement in museum visits. This work will briefly introduce these modules or agents:

- **Visual Retrieval Pipeline:** A three-stage image processing workflow for artwork recognition, including object detection, semantic embedding retrieval, and local feature reranking.
- **Automated PoI Annotation:** A cross-modal pipeline that identifies and annotates Points of Interest (PoIs) within artworks leveraging both pre-existing audio guide content and vision-language processing.
- **Personalized Multimodal Interaction:** A dynamic prompt engineering framework that tailors responses based on user profiles and interaction context, enabling chat, voice, and image-based input.
- **Ethical Guardrails and Safety Filters:** Mechanisms to ensure content appropriateness, cultural alignment, and responsible AI behavior in public deployment.

In Sect. 1 we introduce the motivation and scope of our work, situating the VisualRAG + PoI + Speech framework within the broader field of HCI in cultural heritage. Section 2 reviews prior research on multimodal museum interfaces,

RAG (Retrieval Augmented Generation) in visual applications, content understanding and generation for heritage collections, and ethical considerations in AI-mediated museum systems. In Sect. 3, we present our proposed framework, detailing the multi-agentic flow and the individual modules for VisualRAG, Point-of-Interest Detection, the Speech Interface, and the Personalization and Ethics mechanisms. Section 4 describes our experimental evaluation plan and metrics, performance analysis, user experience findings, and expert feedback with ethical reflections. Finally, Sect. 5 concludes with a summary of contributions, discusses limitations, and outlines directions for future work.

2 Related Work

2.1 Multimodality in HCI and Museum Settings

Recent advancements in multimodal HCI for cultural heritage emphasize the integration of visual, auditory, haptic, and gestural modalities to enrich visitor experiences, supporting naturalistic exploration and multisensory learning [2]. Mobile and wearable museum guides now leverage voice, touch, and group-synchronized audio to facilitate inclusive tours, accommodating diverse visitor needs such as visually impaired users and group-based experiences [15,17]. Virtual tour systems like VirtuWander demonstrate how coupling gesture, voice, and language-model-driven conversational guidance can personalize navigation and engagement in both physical and simulated museum contexts [21]. Empirical evaluations indicate that such multimodal approaches significantly improve visitor satisfaction, learning outcomes, and emotional engagement compared to traditional single-modality guides [9].

More broadly, advances in multimodal audio–language models for emotion and intent recognition offer promising avenues for museum guides to respond empathetically to visitor affective states. For example, Bellver et al. [3] present a Multimodal Audio–Language Model that fuses audio encoding and language understanding to categorize speech emotions, illustrating how such models can enrich interactive systems by tailoring responses not only to content but also to visitor mood and engagement levels. Similarly, gamified participatory sensing approaches influence tourist behavior and satisfaction through adaptive incentive mechanisms and real-time context collection [12], and on-site trip planning support systems leverage dynamic information on tourism spots to curate personalized recommendations in situ [10].

2.2 RAG in Visual Applications and Cultural Heritage

Retrieval-Augmented Generation (RAG) has been extended to vision tasks by coupling image embeddings with external knowledge sources, improving contextual understanding and generation in domains such as medical imaging and multimodal question answering [23]. Systems like FolkRAG apply RAG to cultural heritage archives by integrating domain-specific retrieval pipelines that fetch metadata and historical context, grounding model outputs in authoritative records and enabling intuitive archival exploration [13].

2.3 Content Understanding and Generation in Cultural Heritage

Visual Question Answering (VQA) frameworks allow visitors to pose natural language questions about exhibits and receive contextually accurate responses, bridging the gap between raw imagery and interpretive metadata in museum settings [4]. Recent work investigates using GPT-3 to generate artwork descriptions at runtime, avoiding costly manual annotation and enabling on-the-fly VQA and captioning for cultural heritage collections [5]. Recent work [16] explore generation of QA (Question-Answer) pairs from visual information and curated text context in art and museum QA systems.

3 Proposed Framework

Our unified framework presents an advanced human–computer interface tailored for cultural heritage sites, combining VisualRAG, Point-of-Interest Detection, Speech Interaction. We design an integrated pipeline that fluidly transitions between visual, spatial, and conversational channels. This approach supports visitors as active co-creators of their museum experiences, allowing the system to respond naturally different possible inputs: images, location context, and spoken queries without forcing rigid interaction patterns.

3.1 Multi-agentic Flow

In our multi-agentic design, each module (Artwork Detection through vision, Point of Interest detection, Speech interface, Persona profile manager, and Content Databases) operates as a semi-autonomous agent that communicates via well-defined message channels (see Fig. 1 for a detailed diagram of the architecture). The Vision agent ingests visitor images and identifies artworks; the PoI agent signals important keypoints within exhibit selections; the Guide agent backed by the LLM (Large Language Model) subscribes to the user's and other module's feeds to generate narrative responses; and the Speech agent mediates natural language queries and vocalizes replies. A lightweight "Shield" agent oversees content safety and domain relevance, ensuring that every response complies with museum guidelines before the Guide module provides it to the visitor as text or audio. By decoupling responsibilities across agents, the framework achieves both robustness and extensibility: new capabilities (e.g. gesture recognition or Augmented Reality overlays) can be added as separate agents without reworking the entire pipeline.

Modeling a museum visit as a collaboration among specialized agents reflects the inherently social, exploratory nature of cultural heritage experiences. Visitors encounter artworks, move through spaces, and ask questions in unpredictable sequences; a monolithic system would struggle to adapt fluidly to these dynamics. In contrast, multi-agentic flow enables context-aware coordination: as a visitor gazes at a sculpture, the user agent triggers a VisualRAG request, the Guide agent retrieves and fuses relevant content, and the Speech agent offers it in the

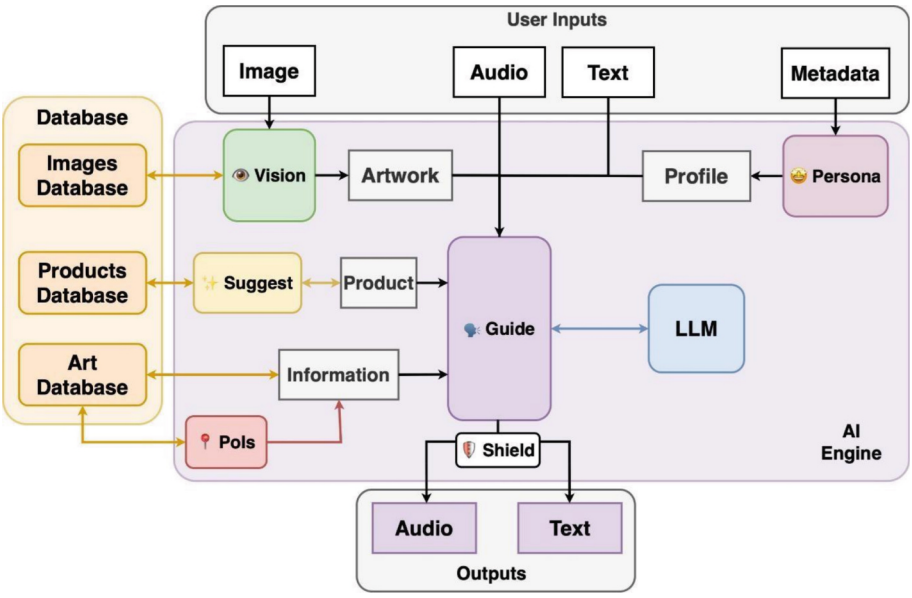


Fig. 1. Multi-agentic architecture of the VisualRAG + PoI + Speech framework.

visitor's preferred modality (Text or Speech). This agent-based orchestration not only mirrors human guide teamwork, but also provides a methodological blueprint for building scalable, adaptable HCI systems in any complex, multi-modal environment.

3.2 VisualRAG

VisualRAG extends retrieval-augmented generation to images captured in situ by museum visitors. When a visitor frames an artwork or artifact, the system extracts a compact visual representation and uses it to retrieve relevant textual passages from a curated heritage knowledge base. These passages are then fused with the visual features to generate a concise narrative tailored to the specific object and visitor context.

VisualRAG is realized as a three-stage pipeline that transforms a visitor's photo of an artwork into a rich, context-aware narrative. In the first *Detection* stage, the system identifies and localizes each distinct object or artwork within the input image, cropping individual elements for subsequent processing. Next, during the *Retrieval* stage, the cropped visuals are matched against a curated vector database of artwork embeddings to fetch the top-K candidate records along with their associated textual descriptions and historical context. Finally, in the *Reranking* stage, lightweight local feature comparisons are applied to refine the initial matches, reordering the candidate list based on geometric and semantic alignment before passing the enriched context to the language

generation component. This modular design allows each step to be optimized independently, ensuring both responsiveness and accuracy in real-time museum settings.

In practice, this spatial awareness enables hands-free guidance through large or complex exhibitions. Visitors need not navigate menus or scan QR codes; instead, the system gently suggests content as they approach each artwork. Such ambient assistance encourages serendipitous discovery and supports varied exploration styles, from goal-oriented tours to free-form wandering.

3.3 Point-of-Interest Detection

The PoI system begins by leveraging vision-language models to detect predefined categories. such as people, objects, places, and landscapes, by analyzing both image content and associated textual sources (e.g., audio-guide transcripts). Advanced AI then generates detailed captions for each detected element, building a structured inventory of visual features. A subsequent multimodal refinement step uses LLMs to align these visual detections with curated museum text, removing spurious or non-significant annotations. Finally, precise bounding box coordinates for each PoI are stored in a reusable graph database, allowing new exhibitions or temporary collections to be integrated seamlessly without altering the core schema and supporting flexible content updates and curator oversight.

The Content Management System (CMS) for Points of Interest serves as the curator's primary interface for defining, organizing, and enriching exhibit metadata that underlies the PoI Detection pipeline (Fig. 2 showcases CMS interface). Through a web-based dashboard, curators can create PoI records by uploading high-resolution images, automatically and manually annotating key visual elements, and linking each record to textual descriptions, audio clips, and related multimedia assets. A flexible schema allows custom fields (such as artist biography, provenance notes, or thematic tags) and points of interest to be added without altering the core database structure, enabling rapid adaptation for special exhibitions or rotating galleries. As new exhibits are introduced, the CMS automatically curates content, keeping it reviewable and editable.

3.4 Multimodal Interface

Our Speech Interface unifies ASR (Automatic Speech Recognition) and TTS (Text-to-Speech) into a conversational layer that respects individual interaction preferences. Visitors can speak naturally (asking "What is this?" or "Tell me more about the artist") and the system will ground responses in both the current visual context and the exhibit's metadata. Dialogue state persists across turns, allowing follow-up questions ("And who inspired her?") without repeated qualifiers (Fig. 3).

By offering speech as an alternative to text, the interface increases accessibility for users with limited difficulties to type or visual impairments. Moreover, visitors may choose to combine modalities communicate with text, then ask

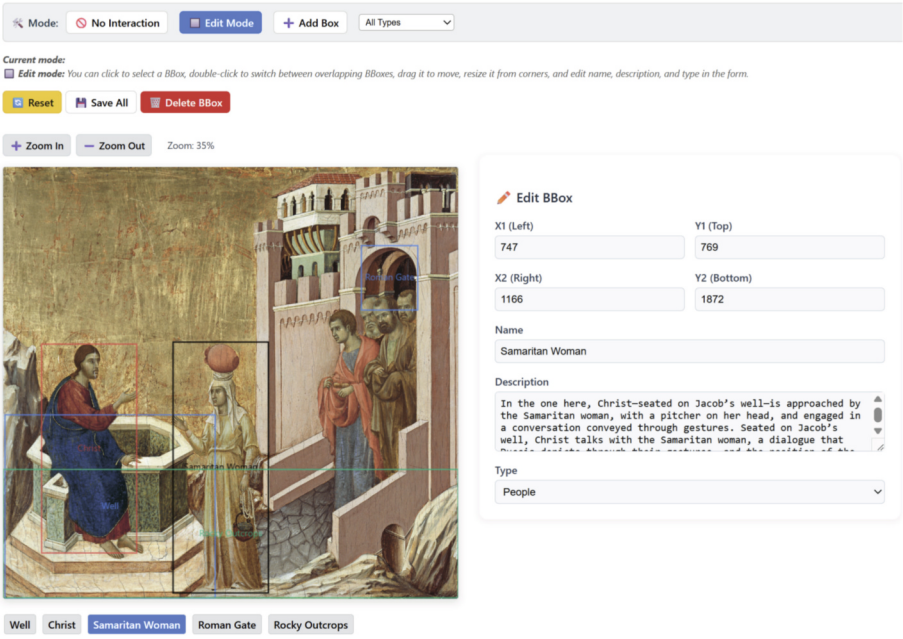


Fig. 2. Bounding Box CMS interface for Point of Interest modification.

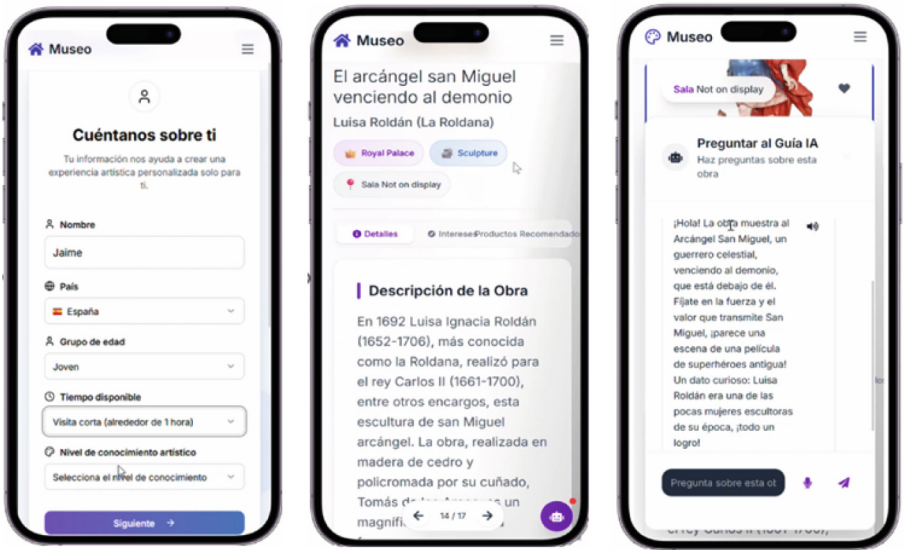


Fig. 3. Screenshots of the designed UX (User Experience). On the left: the onboarding form. On the center: Factual information about the artwork. On the right: Interactive information with the AI assistant.

a question or switch entirely to voice, reinforcing the system's adaptability to personal comfort and situational constraints.

3.5 Personalization

Personalization weaves through every stage of the interaction, shaping content and modality to individual visitor profiles and real-time behavior. Initial preferences, such as language, depth of detail, or interest areas (e.g., art history versus conservation), are set at the start, as well as visit duration (to dynamically accommodate the answers length to the time remaining). As the visit unfolds, the system dynamically adjusts narrative complexity based on the user expertise levels, recommending simpler overviews for casual browsers or deep-dive lectures for enthusiasts.

3.6 Ethics

This research prototype has been guided by key ethical principles, including human agency, privacy, fairness, and transparency. While not yet deployed with real users, the system incorporates early safeguards and design choices aligned with responsible AI development.

Human oversight is enabled through a content management interface allowing curators to review and edit automatically generated outputs. The user interface also supports basic control over interaction flow, preserving user autonomy. In addition, visitors may opt out of the AI-based system entirely and access only the standard textual information available on the museum website.

Privacy is addressed by collecting only minimal, non-identifiable data (e.g., age range, country, art knowledge). Uploaded images are processed in memory without storage, and all communications are encrypted. As this prototype is used internally, formal data audits are not yet implemented but are planned for later stages.

Fairness and inclusivity were considered when designing personalization features for diverse age groups and knowledge levels. Basic content filtering is employed to prevent harmful outputs, though no formal bias audits have yet been conducted.

Transparency and accountability are supported through clear usage disclaimers and internal logging. The system architecture and decision logic are documented, with future versions expected to include explainability tools and formal governance mechanisms.

This approach reflects an "ethics by design" mindset, acknowledging current limitations while laying the groundwork for responsible future deployment in public cultural settings.

4 Experimental Evaluation

4.1 Evaluation Methodology

The evaluation of took place in a focus group on June 12, 2025, with 30 participants, including 12 expert panelists (directors, digital-strategy leaders, and

innovation experts). Table 1 shows the diverse profiles. There was representation from several Spanish cultural heritage institutions of various domains (mainly art museums). The implemented features were showcased to the participants through a presentation and videos of the system's prototype, and finally provided their feedback on-site or after the meeting through a questionnaire.

Table 1. Focus Group Participant Distribution

Institution Type	Count	Role Category
Art Museums	8	Directors, Curators
Cultural Heritage Sites	6	Digital Strategy Leaders
Religious Monuments	4	Innovation Experts
Sports/Entertainment	2	Technology Managers
Project Experts	10	Technical Evaluators
Total Participants	**30**	**Mixed Expertise**

Through a focus group interview, the evaluation focused on the system's ability to demonstrate the flexibility and potential of our technologies in cultural settings, serving as a foundation for future development within the domain.

4.2 Dataset and Technical Implementation

The evaluation dataset comprises 1,603 artworks and artists automatically crawled from various museum websites via a UPM-developed pipeline, with a curated subset of 20 emblematic works (spanning Renaissance to contemporary periods) selected to minimize LLM hallucinations, ensure the graph database was manageable and reviewable, and ease up testing of the different functionalities developed. Table 2 showcases this test set.

Table 2. Evaluation Dataset Characteristics

Museum	Artworks Selected	Art Periods Covered
Museum1	10	Renaissance to Modern
Museum2	10	Medieval to Contemporary
Total Dataset	**20**	**Multi-period**

4.3 System Performance Analysis

The technical performance evaluation measured response time, accuracy, and scalability under varying loads: the three-stage Visual RAG pipeline (detection,

retrieval, reranking) consistently met real-time targets across diverse museum environments and lighting conditions while delivering high artwork identification accuracy. The PoI detection module reduced manual curator review time significantly, all without sacrificing precision across artworks from multiple periods and styles (Table 3).

Table 3. System Performance Metrics Across Core Functions

Function	Processing Time	Efficiency Gain
VisualRAG (20 image DB)	<1 s	Real-time (>30 fps)
PoI Detection	Variable	Automation
Personalization	<1 s	Adaptive

The speech technology components demonstrated effective performance in museum environments. The TTS synthesis maintained high audio quality at 16 kHz and 32 kbps, with user preference options for voice gender selection across supported languages. The other developed components automate processes traditionally costly.

4.4 User Experience

The personalization framework dynamically integrates visitor characteristics (age, visit duration, and art proficiency). Modeling its system prompts to tailor both language complexity (from storytelling for children to technical analysis for experts) and response length (concise facts for quick visits versus detailed explanations for longer stays), the system automatically adjusts terminology, explanation depth, and cultural references to match each user's expertise and preferences.

4.5 Expert Feedback and Ethical Considerations

The focus group of 30 museum professionals (including directors, digital strategy leaders, and innovation experts) highlighted the potential to boost visitor engagement while preserving educational integrity and cultural authenticity. Participants praised the automated PoI detection for drastically reducing manual curation effort yet maintaining curator oversight, and valued the system's ability to handle new or temporary collections with minimal intervention. They also commended the multimodal interfaces (speech recognition and TTS) for enhancing accessibility and maintaining conversational context across modes, catering to diverse visitor needs.

Ethical and safety features were deemed essential for public deployment: the Shield Module's content moderation, visible disclaimers, and minimal data collection policies aligned with institutional requirements for responsible AI in

cultural heritage. This feedback underlines the need for robust moderation and privacy safeguards to deliver both innovative and trustworthy museum experiences.

5 Conclusion and Future Work

The system's key innovations include context-aware personalization, which dynamically adapts responses based on validated user profiles, and automatic detection of Points of Interest (PoIs) in artworks, achieved through the sophisticated combination of vision–language models and curated museum content. Furthermore, the system incorporates multimodal Retrieval-Augmented Generation (RAG)-based image recognition, providing quick contextual information. The multimodal AI assistant interaction, supporting speech, text, and image-based inputs and outputs, significantly enhances accessibility and engagement for diverse audiences.

The comprehensive development effort undertaken by the UPM team involved several critical tasks. This included the meticulous pre-processing of content data, featuring automatic transcription of audio guides using high-performing ASR models. The system also performs automatic extraction of visual and semantic descriptions from artwork images through a multi-stage pipeline powered by advanced vision–language models. Additionally, the project involved the development of a synthetic dataset for in-context learning, generating tailored Q&A dialogues for various user profiles, and the implementation of a custom Content Management System (CMS) to allow curators to review and edit PoIs. The design and deployment of a responsive UX frontend further enabled intuitive user interaction with both the backend system and the physical artworks.

From an ethical and responsible AI perspective, the system incorporates safeguards inspired by the EIC ASTOUND Pathfinder project's work (101071191) on transparency, safety, and user trust. This demonstrates a foundational design principle, where ethical considerations were integrated from the outset rather than treated as an afterthought. These safeguards include content moderation and filtering through LlamaGuard v3.0, visible disclaimer alerts to inform users they are interacting with an AI system, and minimal data collection (e.g., age groups and general preferences instead of personal identifiers). Furthermore, HTTPS-encrypted communication, local logging mechanisms, cautious and reduced usage of images (artworks only, no personal photos), and manual validation of responses in potentially sensitive scenarios were incorporated to ensure privacy and reliability. This proactive integration of ethical considerations sets an example for "ethics by design," advocating for responsible AI awareness in the design of this prototype and laying the groundwork for its long-term deployment in sensitive public domains like museums.

Finally, the presentation of the demo during the Focus Group on June 12, 2025, and the valuable feedback received from leading Spanish museum representatives affirm the system's potential to revolutionize visitor experiences. This feedback will serve as a crucial guide for upcoming enhancements.

5.1 Upcoming Features and Improvements

Key planned features and improvements for implementation and evaluation in the upcoming months as part of the continued ASTOUND project efforts include:

- **Advanced Attention Analysis and User Modeling:** The project aims to incorporate additional audio-derived features to predict and analyze user interest and contextual awareness.
- **Dynamic User Modeling:** Integrating sophisticated reasoning modules with interaction-derived data to infer user states, such as attention and engagement, and adapt system behavior accordingly.[1]
- **Refinement of Fine-Tuned Language Models:** A synthetic Q&A and dialogue dataset was created to fine-tune a pre-trained model on museum-specific content, aiming to reduce hallucinations. While this process provided valuable insights, hyperparameter optimization proved more time-consuming than anticipated, and the fine-tuned model's performance did not consistently exceed that of proprietary models. Future work will explore minimizing reliance on external LLMs in favor of local models to enhance privacy.
- **Platform Scalability and Robustness Improvements:** Enhancements will focus on supporting concurrent access by multiple users, robust session and turn management, and the exploration of invisible watermarking techniques to detect and trace generated responses; bolstering transparency and accountability.

Acknowledgments. This work is supported by the European Commission through Projects ASTOUND (101071191—HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) and ASSIST (101201944 - HORIZON-EIC-2024-BOOSTER-IBA-01). In addition, it is supported by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.1303-9/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union".

References

1. Al Shammari, T., Eiman, D.: My heritage companion: an ai-driven mobile experience for visual storytelling. In: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (2025)
2. Azofeifa, J.D., Noguez, J., Ruiz, S., Molina-Espinosa, J.M., Magana, A.J., Benes, B.: Systematic review of multimodal human-computer interaction. Informatics **9**(1), 13 (2022). https://doi.org/10.3390/informatics9010013
3. Bellver, J., Martín-Fernández, I., Bravo-Pacheco, J.M., Esteban, S., Fernández-Martínez, F., D'Haro, L.F.: Multimodal audio-language model for speech emotion recognition. In: Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2024), pp. 288–295 (2024). https://doi.org/10.21437/odyssey.2024-41

[1] For this prototype, uploaded images or audio are processed in-memory and not stored. Logs are stored with a hash identifier, and only non-sensitive personal information is collected. Although this is a preliminary prototype, the system is intended to comply with GDPR and includes opt-in consent and disclaimers.

4. Bongini, P., Becattini, F., Bagdanov, A.D., Del Bimbo, A.: Visual question answering for cultural heritage. In: IOP Conference Series: Materials Science and Engineering, vol. 949, p. 012074. IOP Publishing (2020)
5. Bongini, P., Becattini, F., Del Bimbo, A.: Is gpt-3 all you need for visual question answering in cultural heritage? In: Computer Vision – ECCV 2022 Workshops, Lecture Notes in Computer Science, vol. 13801, pp. 268–281. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25056-9_18
6. Caramiaux, B.: Ai with museums and cultural heritage. In: AI in Museums, pp. 117–130. Springer, Cham (2023)
7. Egbariya, B., Dror, R., Kuflik, T., Shimshoni, I.: Image-based poi identification for mobile museum guides: design, implementation, and user evaluation. Heritage **8**(7) (2025). https://doi.org/10.3390/heritage8070266, https://www.mdpi.com/2571-9408/8/7/266
8. Ferrara, C., Cerquetti, M.: Towards user-centred and context-open museums: a review of mobile apps adopted by the Italian national museum with special autonomy status. J. Cult. Heritage Manage. Sustain. Dev. (2025)
9. Ferrato, A.: Integrating indoor positioning, recommendation, and group synchronization in intelligent museum guides. In: Proceedings of the 2023 ACM Conference on Ubiquitous Computing, pp. 215–224 (2023). https://doi.org/10.1145/3699682.3727572
10. Hidaka, M., et al.: On-site trip planning support system based on dynamic information on tourism spots. Smart Cities **3**(2), 212–231 (2020). https://doi.org/10.3390/smartcities3020013
11. Ivanov, R.: Advanced visitor profiling for personalized museum experiences using telemetry-driven smart badges. Electronics **13**(20), 3977 (2024). https://doi.org/10.3390/electronics13203977
12. Kawanaka, S., Matsuda, Y., Suwa, H., Fujimoto, M., Arakawa, Y., Yasumoto, K.: Gamified participatory sensing in tourism: an experimental study of the effects on tourist behavior and satisfaction. Smart Cities **3**(3), 736–757 (2020). https://doi.org/10.3390/smartcities3030037
13. Kelly, P., Schild, J., Jafari, A.: Folkrag: a retrieval-augmented generation system for cultural heritage materials. Neural Comput. Appl. (2025). https://doi.org/10.1007/s00521-025-11455-4
14. Lushnikova, A., Morse, C., Doublet, S., Koenig, V., Bongard-Blanchy, K.: Self-determination theory applied to museum website experiences: fulfill visitor needs, increase motivation, and promote engagement. In: Proceedings of the European Conference on Cognitive Ergonomics 2023. ECCE 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3605655.3605658, https://doi.org/10.1145/3605655.3605658
15. Mohlin, F.: Implementation and evaluation of a mobile tour guide with group synchronized audio: Using web technologies, QR codes, and speech synthesis. In: Proceedings of the 2024 International Conference on Mobile and Ubiquitous Multimedia (2024)
16. Ramos-Varela, S., Bellver-Soler, J., Estecha-Garitagoitia, M., D'Haro, L.F.: Context or retrieval? evaluating rag methods for art and museum QA system. In: Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pp. 129–136. Association for Computational Linguistics, Bilbao, Spain (2025). https://doi.org/10.18653/v1/2025.iwsds-1.10, https://aclanthology.org/2025.iwsds-1.10/
17. Suikkila, J.: Museum wayfinding with ai-spoken voice guidance. In: Proceedings of the 2025 International Conference on Intelligent User Interfaces (2025)

18. Trichopoulos, G., Konstantakis, M., Caridakis, G., Katifori, A., Koukouli, M.: Crafting a museum guide using chatgpt4. Big Data Cogn. Comput. **7**(3) (2023). https://doi.org/10.3390/bdcc7030148, https://www.mdpi.com/2504-2289/7/3/148
19. Hsieh, T-J., Su, Y-H., Lin, L.S.: Augmented reality art museum mobile guide for enhancing user experience. IEEE Comput. Graph. Appl. **45**(1), 10–20 (2025). https://doi.org/10.1109/MCG.2025.3529981
20. Varga, M.N., Bazazian, D.: Interactive heritage site mobile application on artworks. In: Advances in Representation: New AI- and XR-Driven Transdisciplinarity, pp. 125–139. Springer, Cham (2024)
21. Wang, Z., Yuan, L.P., Wang, L., Jiang, B., Zeng, W.: Virtuwander: enhancing multi-modal interaction for virtual tour guidance through large language models. arXiv preprint arXiv:2401.11923 (2024). https://arxiv.org/abs/2401.11923
22. Li, Y., Ma1, Y., F.L.H.B.: Empowering museum visitor experiences through ai-driven smart guide systems: a model for intelligent museum services. GAS J. Eng. Technol. (GASJET) (2025)
23. Zheng, X., et al.: Retrieval augmented generation and understanding in vision: a survey and new outlook. arXiv preprint arXiv:2503.18016 (2025)

Geometric Goodness-of-Fit Measure for Interval-Valued Data

Dylan Benavides[1(✉)], Oldemar Rodríguez[2], and Maikol Solís[2]

[1] National High Technology Center, San José 10109, Costa Rica
dbenavides@cenat.ac.cr
[2] University of Costa Rica, San José 11501-2060, Costa Rica
{oldemar.rodriguez,maikol.solis}@ucr.ac.cr
https://www.cenat.ac.cr/es/ , https://www.ucr.ac.cr

Abstract. Geometric goodness-of-fit measures have demonstrated strong potential to uncover structural patterns in classical data where each observation is a single point, by quantifying spatial relationships within clouds of points. Building on this foundation, Solís and Hernández (2023) introduced an index based on the area difference between the alpha shape of a two-dimensional projection of the data and its minimal bounding rectangle. In this study, we propose an extension of this geometric index to regression models involving symbolic data represented by intervals. Our approach analyzes rectangular clusters defined by the interval values of predictor and response variables, adapting the alpha shape methodology to address both the central tendency and variability inherent in interval observations.

We introduce an interval geometric fit function that integrates alpha shapes built from the centers and ranges (In this paper, we use the term *range* to refer to the half-range of an interval, i.e., half the width of the interval.) of interval-valued data. This function, defined over the scale parameter α, quantifies geometric goodness-of-fit by revealing both concordance and divergence between the location and variability components. The method is implemented in `R` and applied to benchmark symbolic datasets, demonstrating its ability to reveal geometric structures and provide insights that complement classical symbolic regression metrics such as the coefficient of determination. All developed functions are publicly available and designed for general use in symbolic data analysis.

Keywords: Goodness-of-fit · Interval-valued data · Rectangular clusters · Alpha shapes · Empty space loss function

1 Introduction

Statistical and data analysis methods have traditionally been developed for variables taking single numerical values. However, in many practical situations, this simplification leads to the loss of relevant information. For quantitative variables,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 27–44, 2026.
https://doi.org/10.1007/978-3-032-20322-9_3

a more complete representation can be achieved by modeling data as interval-valued data, where each variable is expressed as a closed interval $[a, b]$, with $a, b \in \mathbb{R}$ and $a \leq b$.

This type of data representation offers computational advantages by summarizing large datasets into more manageable forms while preserving essential information about central tendency and variability. Several methods have been proposed for analyzing interval-valued data, including regression models and dimensionality reduction techniques.

A commonly used measure for assessing goodness-of-fit in symbolic linear regression models based on the centers and ranges method is the symbolic coefficient of determination. This coefficient is calculated by applying separate regression models to the centers and the ranges matrices [8]. Nevertheless, it inherits limitations from the classical R^2, such as increasing when irrelevant predictors are added, tending to be larger in small samples, and ignoring the geometric structure of the data [4].

In the context of classical point-valued data, Solís and Hernández [6] introduced a geometric goodness-of-fit index that measures the loss of empty space between the alpha shape fitted to a point cloud and its minimal bounding rectangle. This index effectively uncovers geometric patterns relating predictor and response variables that traditional measures overlook. Furthermore, Solís and Hernández [9] applied this geometric approach to problems of dimensionality reduction, demonstrating its broader applicability.

Building upon this geometric perspective, the main contribution of this paper is to extend the geometric goodness-of-fit index to symbolic interval-valued data. Instead of analyzing isolated points, we consider rectangular clusters formed by the intervals of predictor and response variables. This extension enables the detection of geometric structures in symbolic regression models beyond what the symbolic coefficient of determination can reveal, providing a richer and more informative assessment of model fit.

2 Theoretical Framework

2.1 Alpha Shapes

Alpha shapes are a continuous generalization of data point clouds [6]. The following is a summary of their construction; for further details, we refer the reader to [5].

- **Voronoi region.** Let S be a finite set of points in $\mathbb{R}^2$. For each point $s \in S$, the Voronoi region of s is defined as:

$$V_s = \left\{ u \in \mathbb{R}^2 \mid \|u - s\|_2 \leq \|u - t\|_2, \forall t \in S \right\}.$$

  The collection of all Voronoi regions is called the Voronoi diagram of S.

- **Delaunay triangulation.** Based on the Voronoi diagram, the Delaunay triangulation is obtained by connecting two points with a line segment if their corresponding Voronoi regions share a common boundary.

- **Alpha shape of S.** For each $s \in S$, let $D_s(\alpha)$ denote the closed disk of radius α centered at s, and define the region $R_s(\alpha) = D_s(\alpha) \cap V_s$. Then, the union
$$U_S(\alpha) = \bigcup_{s \in S} R_s(\alpha)$$
forms a subset of the Voronoi diagram of S. Next, two points are connected by a line segment if they share a boundary within the set $U_S(\alpha)$, generating a new triangulation denoted by $A_S(\alpha)$. The set of triangles $A_S(\alpha)$ is called the alpha complex of S, and the union of these complexes as α increases defines the alpha shape of S.

2.2 Measuring Distances in Point Patterns

Traditionally, correlation has been employed as a statistical tool to assess whether points in a point pattern are independently distributed or exhibit some form of interdependence. However, statistical correlation alone is insufficient to characterize the nature of such dependence or to infer causality.

To enrich the analysis, one can examine the point pattern by measuring spatial spacing and minimal interpoint distances within the dataset [2]. Various types of distances can be quantified in a point pattern, including the following:

- **Pairwise distances.** $d_{ij} = \|x_i - x_j\|$ between all pairs of points x_i and x_j with $i \neq j$ in the pattern.
- **Nearest-neighbour distances.** $d_i = \min_{i \neq j} d_{ij}$, the distance from a data point x_i to the nearest other data point.
- **Empty-space distance.** $d(u) = \min_j \|u - x_j\|$, the distance from a fixed location u to the nearest data point.

A point process is said to be stationary when the statistical properties of any subset of the data are independent of the particular subset chosen [2]. Under this assumption, for a dataset A, the cumulative distribution function of the empty space distance is expressed as $F(r) = P\{d(u) \leq r\}$

The point u represents an arbitrary location within the window containing the dataset, and r is a value greater than or equal to zero. The values of F correspond to probabilities that increase as r grows. Therefore, for this type of process, F is always differentiable.

On the other hand, under the assumption of a stationary process and considering a data set A, the cumulative distribution function corresponding to the neighbor distance is defined as $G(r) = P\{d_i \leq r\}$.

The values of G represent probabilities that increase or remain constant as r grows. Due to the stationarity of the process, this definition is independent of the specific location at which the measurement is taken. Generally, G is not a differentiable function and, therefore, may lack an associated probability density.

2.3 Complete Spacial Randomness Process and Intensity

According to [2] a point process considered to be completely random is characterized by two fundamental properties. First, homogeneity, which means that points exhibit no preference for any particular spatial location. Second, independence, which implies that information obtained in one region of space does not influence or condition outcomes in other regions. This process is known as the homogeneous Poisson point process or complete spatial randomness (CSR).

The homogeneity assumption states that the expected number of points within a specific subregion B of the observation window W is proportional to the area of that subregion, $|B|$, with the proportionality constant being the intensity λ of the process.

The independence assumption imposes an ordering condition that ensures the probability of two points coinciding exactly is negligible.

Complete spatial randomness (CSR) processes are commonly used as null models in statistics, serving as a baseline for comparison against more complex models.

In the analysis of point patterns, the intensity λ represents a fundamental characteristic, defined as the expected number of points per unit area. Its empirical estimate is obtained by dividing the total number of points $n(S)$ by the area of the observation window $|W|$.

Furthermore, since in a Poisson process the variance equals the mean, additional inferences can be made based on this property.

2.4 Regression Models for Interval-Type Symbolic Data

Symbolic data analysis allows modeling complex objects such as intervals, histograms, or distributions. This work focuses on interval-valued data, where each observation is described by a lower and an upper bound. Traditional regression techniques have been extended to this data type mainly through three approaches: the linear models known as the method of centers (CM) and the method of centers and ranges (CRM), as well as nonlinear models (NLM).

Given a dataset $E = \{e_1, \ldots, e_n\}$ consisting of n observations and $m + 1$ interval-valued variables $(X_1, \ldots, X_m, Y)$, each unit e_i (where $i = 1, \ldots, n$) is represented by intervals $x_{ij} = [a_{ij}, b_{ij}]$ for the predictors and $y_i = [y_{Li}, y_{Ui}]$ for the response, with $j = 1, \ldots, m$ indexing the predictor variables.

The CM model replaces each interval by its midpoint. Lima Neto and De Carvalho [8] proposed estimating Y by fitting two independent classical regression models separately to the lower and upper bounds.

$$y_{Li} = \beta_0 + \beta_1 a_{i1} + \cdots + \beta_m a_{im} + \varepsilon_{Li},$$
$$y_{Ui} = \beta_0 + \beta_1 b_{i1} + \cdots + \beta_m b_{im} + \varepsilon_{Ui}.$$

where $\beta_0, \beta_1, \ldots, \beta_m$ are the regression coefficients to be estimated, and ε_{Li} and ε_{Ui} denote the error terms associated with the lower and upper bounds of the i-th observation, respectively.

The CRM model considers both the center and the range of each interval, thereby accounting for the central tendency and variability of the data. Lima Neto and De Carvalho [8] proposed estimating Y by fitting regression models that simultaneously incorporate both components. For interval-valued variables, the centers and half ranges are defined as:

$$x_{ij}^c = \frac{a_{ij} + b_{ij}}{2}, \quad x_{ij}^r = \frac{b_{ij} - a_{ij}}{2}, \quad y_i^c = \frac{y_{Li} + y_{Ui}}{2}, \quad y_i^r = \frac{y_{Ui} - y_{Li}}{2}.$$

Let $x_i^c = (x_{i1}^c, \ldots, x_{im}^c)$ and $x_i^r = (x_{i1}^r, \ldots, x_{im}^r)$ be the vectors of centers and ranges for the predictor variables. The resulting regression models are given by:

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_m^c x_{im}^c + \varepsilon_i^c, \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_m^r x_{im}^r + \varepsilon_i^r. \end{aligned}$$

where $\beta_0^c, \beta_1^c, \ldots, \beta_m^c$ and $\beta_0^r, \beta_1^r, \ldots, \beta_m^r$ are the regression coefficients for the center and range models, respectively. The terms ε_i^c and ε_i^r denote the error terms associated with the center and range regressions for observation.

2.5 Goodness-of-Fit Measures in Symbolic Regression

In symbolic regression, the most commonly used measure of fit is the symbolic coefficient of determination R^2, which quantifies the proportion of variability in the response explained by the model. Under the CRM approach, this assessment is typically performed separately for the centers and ranges of the intervals, resulting in two coefficients: R_c^2 and R_r^2.

These are defined as:

$$R_c^2 = 1 - \frac{\sum_{i=1}^n (y_i^c - \hat{y}_i^c)^2}{\sum_{i=1}^n (y_i^c - \bar{y}^c)^2}, \quad R_r^2 = 1 - \frac{\sum_{i=1}^n (y_i^r - \hat{y}_i^r)^2}{\sum_{i=1}^n (y_i^r - \bar{y}^r)^2},$$

where $\hat{y}_i^c$ and $\hat{y}_i^r$ are the predicted center and range of the response for unit i, and $\bar{y}^c$ and $\bar{y}^r$ are the corresponding sample means.

To obtain an overall measure of fit, three common aggregations of R_c^2 and R_r^2 are considered:

$$\begin{aligned} R_A^2 &= \min(R_c^2, R_r^2), \\ R_B^2 &= \frac{R_c^2 + R_r^2}{2}, \\ R_C^2 &= \max(R_c^2, R_r^2). \end{aligned}$$

These three variants provide conservative (R_A^2), balanced (R_B^2), and optimistic (R_C^2) perspectives on model performance.

Although widely used, these measures focus solely on variance explained and may fail to capture structural or spatial characteristics in the data, especially

in the presence of nonlinear or irregular patterns. This limitation has motivated the development of alternative evaluation criteria that assess model fit from a geometric perspective.

2.6 Geometric Goodness-of-Fit Index on Classical Data

Solís and Hernández [6] define the geometric goodness-of-fit index $R^2_{\text{Geom},j,\alpha}$ as a measure of the proportion of empty space between the alpha shape constructed from the bivariate projection (X_j, Y) and the smallest rectangular window B_j that contains all the data points in that projection. The purpose of the index is to detect geometric structures that may arise from the relationship between the predictor X_j and the response variable Y.

Given a radius parameter $\alpha > 0$, the alpha shape $\mathcal{R}_{j,\alpha}$ is computed using the standard procedure based on Delaunay triangulations (see [5]). The area of this shape, denoted as $\text{Area}(\mathcal{R}_{j,\alpha})$, is compared to the area of the bounding box B_j, denoted as $\text{Area}(B_j)$. The index is then defined as:

$$R^2_{\text{Geom},j,\alpha} = 1 - \frac{\text{Area}(\mathcal{R}_{j,\alpha})}{\text{Area}(B_j)},$$

where $\text{Area}(B_j)$ corresponds to the product of the ranges of X_j and Y in the sample. By construction, this quantity lies in the interval $[0, 1]$, where values close to 1 indicate structured or sparse configurations, and values close to 0 suggest a random or uniform distribution of the data within the domain.

Importantly, the function $f_j : \mathbb{R}^+ \to [0, 1]; \alpha \to R^2_{\text{Geom},j,\alpha}$ is monotonic non-increasing, since the alpha shape grows as α increases, gradually filling more of the domain. The curve $f_j(\alpha) := R^2_{\text{Geom},j,\alpha}$ captures the evolution of the geometric structure at different spatial scales. Based on the shape of these curves, Solís and Hernández identify five qualitative patterns that reveal internal and external characteristics of the data: plateaus, steep descents and asymptotic behaviors, which help distinguish between noise-like configurations and informative spatial patterns [6].

This index provides a geometric complement to traditional statistical measures, especially useful in contexts where structure or irregularity in the spatial arrangement of data is informative, motivating its extension to symbolic regression scenarios involving interval-valued data.

3 Methodology

In this section, we describe the procedure used to construct the geometric goodness-of-fit index for interval-valued data.

3.1 Construction of the Geometric Goodness-of-Fit Index for Interval-Valued Data

Consider an interval matrix X with n rows and m columns of interval-valued predictor variables, and an interval-valued response vector Y of size $n \times 1$.

$$X = \begin{bmatrix} [a_{11}, b_{11}] & [a_{12}, b_{12}] & \cdots & [a_{1\,m}, b_{1\,m}] \\ [a_{21}, b_{21}] & [a_{22}, b_{22}] & \cdots & [a_{2\,m}, b_{2\,m}] \\ \vdots & \vdots & \ddots & \vdots \\ [a_{n1}, b_{n1}] & [a_{n2}, b_{n2}] & \cdots & [a_{nm}, b_{nm}] \end{bmatrix} \quad , \quad Y = \begin{bmatrix} [y_{L1}, y_{U1}] \\ [y_{L2}, y_{U2}] \\ \vdots \\ [y_{Ln}, y_{Un}] \end{bmatrix}$$

From this matrix $X \in \mathbb{IR}^{n\times m}$ we define two new matrices $X^c, X^r \in \mathbb{R}^{n\times m}$, which contain the centers and the ranges of X, respectively. And from the interval-valued response vector $Y \in \mathbb{IR}^{n\times 1}$, we define two new vectors $Y^c, Y^r \in \mathbb{R}^{n\times 1}$, which contain the centers and the ranges of Y, respectively.[1]

We denote C_j with $j = 1, \ldots, m$ the rectangular clustering associated to the variables X_j and Y and is defined by

$$C_j = \{[a_{1j}, b_{1j}] \times [y_{L1}, y_{U1}], \ldots, [a_{nj}, b_{nj}] \times [y_{Ln}, y_{Un}]\} = \left\{R_1^j, \ldots, R_n^j\right\}$$

For each rectangle $R_i^j \in C_j$ with $i = 1, \ldots, n$ we obtain its center, which is given by

$$\left(\frac{a_{ij} + b_{ij}}{2}, \frac{y_{Li} + y_{Ui}}{2}\right)$$

Thus, we define the point cloud of centers associated to the rectangular clustering C_j as

$$N_j^c = \left\{\left(\frac{a_{ij} + b_{ij}}{2}, \frac{y_{Li} + y_{Ui}}{2}\right); i = 1, \ldots, n\right\}$$

Furthermore, we define the smallest rectangular observation window containing the point cloud N_j^c as

$$B_j^c = \left[\min_{i=1,\ldots,n} \frac{a_{ij} + b_{ij}}{2}, \max_{i=1,\ldots,n} \frac{a_{ij} + b_{ij}}{2}\right] \times \left[\min_{i=1,\ldots,n} \frac{y_{Li} + y_{Ui}}{2}, \max_{i=1,\ldots,n} \frac{y_{Li} + y_{Ui}}{2}\right]$$

Let $\mathcal{R}_{j,\alpha}^c$ the alpha shape generated by the pair of variables $\left(X_j^c, Y^c\right)$ for $j = 1 : m$ and of radius α, where X_j^c is the column j of X^c.

Thus, we define the geometric goodness-of-fit index for the centers of the interval-valued data as

$$I_{j,\alpha}^c = 1 - \frac{A\left(\mathcal{R}_{j,\alpha}^c\right)}{A\left(B_j^c\right)} \tag{1}$$

[1] $\mathbb{IR}^{n\times m} := \{A = (A_{ij})_{i=1,\cdots,n;j=1,\cdots,m} : A_{ij} = [a_{ij}, b_{ij}], a_{ij}, b_{ij} \in \mathbb{R}, a_{ij} \leq b_{ij}\}$.

To better understand how the geometric structure evolves across scales, we define the following function:

$$f_j^c : \mathbb{R}^+ \to [0,1]; \alpha \to I_{j,\alpha}^c \quad \text{for} \quad j = 1, \ldots, m$$

This curve captures the progressive filling of the empty space as the parameter α increases. It provides multiscale information about the spatial organization of the centers in the interval-valued data.

On the other hand, for each rectangle $R_i^j \in C_j$ with $i = 1, \ldots, n$, we obtain its ranges, which is given by:

$$\left(\frac{b_{ij} - a_{ij}}{2}, \frac{y_{Ui} - y_{Li}}{2} \right)$$

Thus, we define the point cloud of ranges associated to the rectangular clustering C_j as

$$N_j^r = \left\{ \left(\frac{b_{ij} - a_{ij}}{2}, \frac{y_{Ui} - y_{Li}}{2} \right), i = 1, \ldots, n \right\}$$

Additionally, we define the smallest rectangular observation window containing the point cloud N_j^r as

$$B_j^r = \left[\min_{i=1,\ldots,n} \frac{b_{ij} - a_{ij}}{2}, \max_{i=1,\ldots,n} \frac{b_{ij} - a_{ij}}{2} \right] \times \left[\min_{i=1,\ldots,n} \frac{y_{Ui} - y_{Li}}{2}, \max_{i=1,\ldots,n} \frac{y_{Ui} - y_{Li}}{2} \right]$$

Let $\mathcal{R}_{j,\alpha}^r$ the alpha shape generated by the pair of variables $\left(X_j^r, Y^r\right)$ for $j = 1 : m$ and of radius α, where X_j^r is the column j of X^r.

Thus, we define the geometric goodness-of-fit index for the ranges of the interval-valued data as

$$I_{j,\alpha}^r = 1 - \frac{A\left(\mathcal{R}_{j,\alpha}^r\right)}{A\left(B_j^r\right)} \tag{2}$$

To better understand how the geometric variability evolves across scales, we define the following function:

$$f_j^r : \mathbb{R}^+ \to [0,1]; \alpha \to I_{j,\alpha}^r \quad \text{for} \quad j = 1, \ldots, m$$

This curve reflects how the dispersion structure encoded in the ranges progressively fills the observation window as α increases. It offers multiscale insights into the geometric behavior of the variability component in the interval-valued data.

Now we define the functions $\phi_j, \psi_j : \mathbb{R}^+ \to [0,1]$ with

$$\phi_j(\alpha) = \frac{f_j^c(\alpha) + f_j^r(\alpha)}{2} \ , \ \psi_j = \frac{|f_j^c(\alpha) - f_j^r(\alpha)|}{2} \tag{3}$$

We define for each $j = 1, \ldots, m$ the following **interval geometric fit function** of a real variable

$$F_j : \mathbb{R}^+ \to [0,1];\ \alpha \to [\phi_j(\alpha) - \psi_j(\alpha), \phi_j(\alpha) + \psi_j(\alpha)] \tag{4}$$

Which can also be viewed as

$$F_j(\alpha) = [\min(f_j^c(\alpha), f_j^r(\alpha)), \max(f_j^c(\alpha), f_j^r(\alpha))] \quad \text{for} \quad j = 1, \ldots, m$$

Finally, we introduce the *Relative Area Under the Curve* ($RAUC$) as a normalized metric to quantify the degree of concentration of the distribution. It is defined as

$$RAUC = \frac{AUC}{TA} = \frac{\sum_{i=1}^{n} \phi_{i-1}\,(\alpha_i - \alpha_{i-1})}{\alpha_n - \alpha_0}, \qquad 0 < RAUC < 1, \tag{5}$$

where AUC denotes the area under the curve $\phi(\alpha)$ and TA represents the total area of the rectangular observation window containing the curves $\phi(\alpha) - \psi(\alpha)$ and $\phi(\alpha) + \psi(\alpha)$. The function $\phi(\alpha)$ is given by:

$$\phi(\alpha) = \frac{f^c(\alpha) + f^r(\alpha)}{2} = \begin{cases} \phi_0 = \frac{f^c(\alpha_0)+f^r(\alpha_0)}{2}, & \alpha \in [\alpha_0, \alpha_1), \\ \phi_1 = \frac{f^c(\alpha_1)+f^r(\alpha_1)}{2}, & \alpha \in [\alpha_1, \alpha_2), \\ \vdots & \vdots \\ \phi_n = \frac{f^c(\alpha_n)+f^r(\alpha_n)}{2}, & \alpha \in [\alpha_{n-1}, \alpha_n]. \end{cases}$$

3.2 Interpretation of the Geometric Goodness-of-Fit Index for Interval-Valued Data

The previously defined functions f_j^c and f_j^r capture the evolution of the geometric goodness-of-fit for the centers and ranges, respectively, as the parameter α varies. These functions describe how the empty space within the observation window is progressively filled by the alpha shape, offering multiscale insights into the spatial organization of the data.

When either f_j^c or f_j^r approaches zero as α increases, this suggests a spatial distribution resembling a random configuration within the observation window. Conversely, values closer to one indicate the presence of structured or sparse geometric patterns in the centers or ranges of the intervals.

In particular, under the assumption of complete spatial randomness (CSR) with intensity λ, the complementary empty space distribution function is given by

$$1 - F_{\mathrm{CSR}}(\alpha) = \exp(-\lambda \pi \alpha^2),$$

which serves as a benchmark to assess whether the geometric behavior of the data deviates from randomness.

To integrate the information from both location and variability, the functions ϕ_j and ψ_j (3) summarize the average geometric fit and the discrepancy between the geometric patterns of the centers and ranges, respectively. The interval-valued function F_j, previously introduced in (4), acts as a geometric consistency band. Its width reflects the degree of alignment or divergence between the structural patterns of the two components. A narrow interval implies similar spatial organization between center and range, while a wider interval indicates differences in their geometric behaviors.

On the other hand, the $RAUC$ metric provides a global indicator of how efficiently the intervals saturate the observation space. Values close to $RAUC \approx 1$ reveal a rapid filling process, associated with a concentrated and coherent distribution, whereas values near $RAUC \approx 0$ indicate a slower filling dynamics, characteristic of a more dispersed or heterogeneous configuration.

This approach provides a richer geometric interpretation of interval-valued data, complementing traditional fit measures and facilitating enhanced variable selection and model evaluation in symbolic regression.

3.3 Pseudocode of the Algorithm

Algorithm 1 Computation of the geometric goodness-of-fit index for interval-valued data

Require: Interval-valued data matrix $X \in \mathbb{IR}^{n\times(m+1)}$, where the first m columns represent predictor variables and the last column is the response variable Y.

Ensure: Interval-valued function $F_j : \mathbb{R}^+ \to [0,1]$ for each predictor variable X_j.

1: Compute the center matrix X^c and the range matrix X^r from X.
2: **for** $j = 1$ to m **do**
3: Construct rectangular clustering $C_j = \{[a_{ij}, b_{ij}] \times [y_{Li}, y_{Ui}]\}_{i=1}^{n}$
4: Extract the center point cloud $N_j^c = \left\{ \left(\frac{a_{ij}+b_{ij}}{2}, \frac{y_{Li}+y_{Ui}}{2} \right) \right\}_{i=1}^{n}$
5: Extract the range point cloud $N_j^r = \left\{ \left(\frac{b_{ij}-a_{ij}}{2}, \frac{y_{Ui}-y_{Li}}{2} \right) \right\}_{i=1}^{n}$
6: Determine the minimal bounding rectangles B_j^c and B_j^r for N_j^c and N_j^r, respectively
7: Construct the alpha shapes $\mathcal{R}_{j,\alpha}^c$ and $\mathcal{R}_{j,\alpha}^r$ for a given radius α
8: Compute the areas $A(\mathcal{R}_{j,\alpha}^c)$, $A(\mathcal{R}_{j,\alpha}^r)$, $A(B_j^c)$, and $A(B_j^r)$
9: Evaluate the geometric indices

$$I_{j,\alpha}^c = 1 - \frac{A(\mathcal{R}_{j,\alpha}^c)}{A(B_j^c)}, \quad I_{j,\alpha}^r = 1 - \frac{A(\mathcal{R}_{j,\alpha}^r)}{A(B_j^r)}$$

10: Define the functions $f_j^c(\alpha) = I_{j,\alpha}^c$, $f_j^r(\alpha) = I_{j,\alpha}^r$
11: Define the final interval geometric fit function

$$F_j(\alpha) = \left[\min\left(f_j^c(\alpha), f_j^r(\alpha)\right), \max\left(f_j^c(\alpha), f_j^r(\alpha)\right)\right].$$

12: **end for**

The algorithm above describes the step-by-step procedure for computing the geometric goodness-of-fit index. It integrates both central tendency and variability by analyzing the centers and ranges of rectangular clusters. This approach allows us to capture structural relationships in symbolic data through a geometric perspective.

All computations were performed in R using functions specifically written for this study, based on the existing `RSDA` and `spatgeom` packages.

4 Applications on Real Data Sets

4.1 Experiment 1

In this experiment, the dataset `abalone` of the `RSDA` package was used, which groups a series of observations on abalone characteristics. This symbolic table has 24 observations of 7 interval-valued variables. The variables are: `LENGTH`, `DIAMETER`, `HEIGHT`, `WHOLE_WEIGHT`, `SHUCKED_WEIGHT`, `VISCERA_WEIGHT` and `SHELL_WEIGHT`.

The response variable is `WHOLE_WEIGHT` and the other six variables are using as predictive variables. The results obtained are shown below.

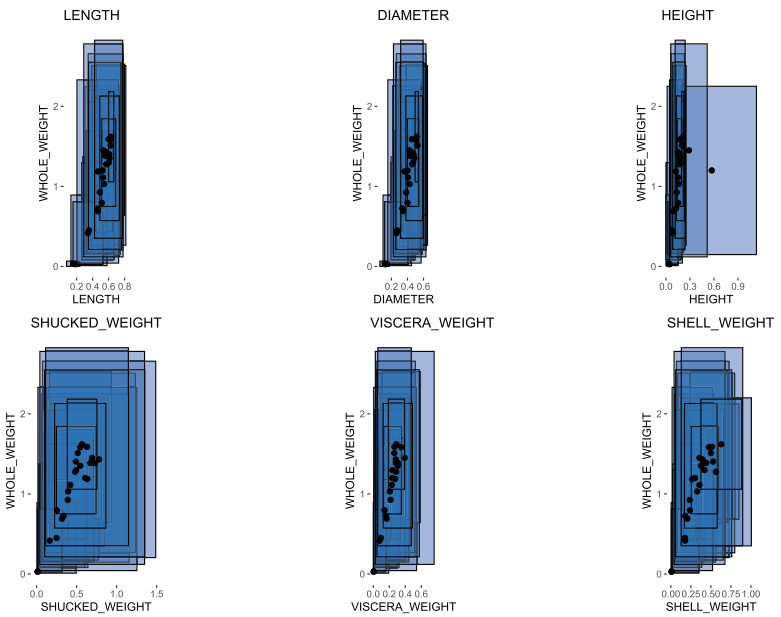


Fig. 1. Rectangular representations of interval ranges for each predictorâĂŞresponse pair in the interval dataset `abalone`, with a point cloud per pair showing the centers of the rectangles (N_j^c).

Figure 1 shows the rectangular groupings formed between each predictor variable and the response variable `WHOLE_WEIGHT`. It can be observed that the

variables LENGTH, DIAMETER, and HEIGHT, on average, exhibit lower variability (smaller intervals) than the other three variables. However, there is one observation in HEIGHT with a range much larger than the average range of this variable, specifically $[0.02, 1.13]$.

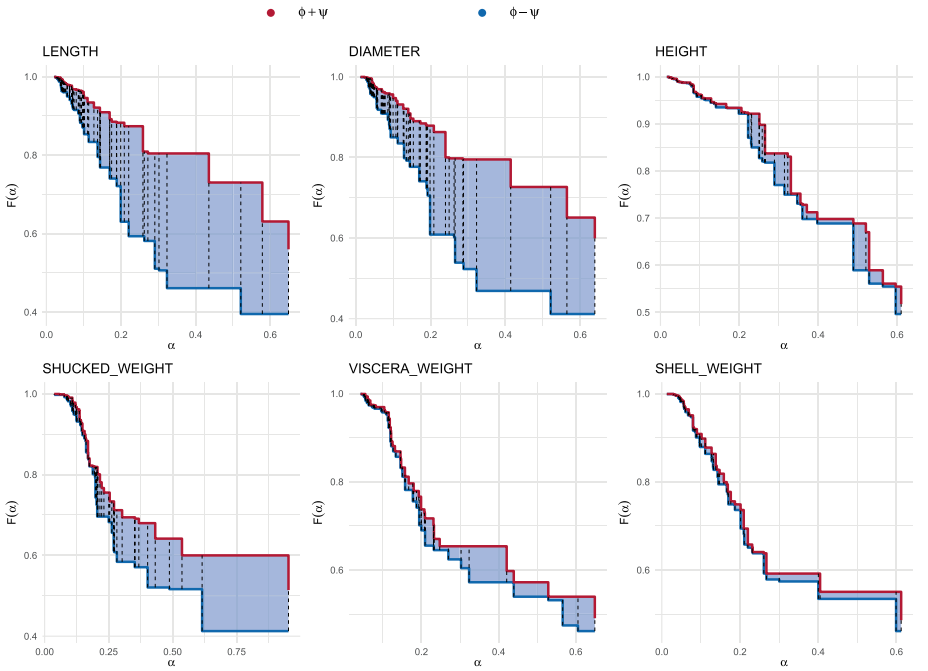


Fig. 2. Interval geometric fit function $F_j(\alpha)$ for each variable j of the abalone dataset.

Figure 2 shows the interval-valued functions $F_j(\alpha)$ for six symbolic variables. The shaded band therefore visualizes, for each α, the range spanned by the geometric indices of the centers and ranges, providing a compact representation of their structural differences across scales. Note that the $\phi - \psi$ and $\phi + \psi$ curves for all variables stabilize at least around 0.4, indicating evidence of geometric patterns between each predictor and the response variable, with variables such as SHELL_WEIGHT, HEIGHT, and VISCERA_WEIGHT exhibiting similar geometric structures of centers and ranges, as the band between the two curves is narrower than the bands of the other variables.

On the other hand, Fig. 3 shows the relative area under the curve ϕ_j, for each variable j. It can be observed that HEIGHT, DIAMETER and LENGTH have the highest $RAUC$ values, indicating that the average curve $\phi_j(\alpha)$ maintains high values for longer before stabilizing. This implies that the void space is filled more gradually, revealing stronger and more persistent geometric patterns. In contrast, SHUCKED_WEIGHT, VISCERA_WEIGHT, and SHELL_WEIGHT exhibit lower

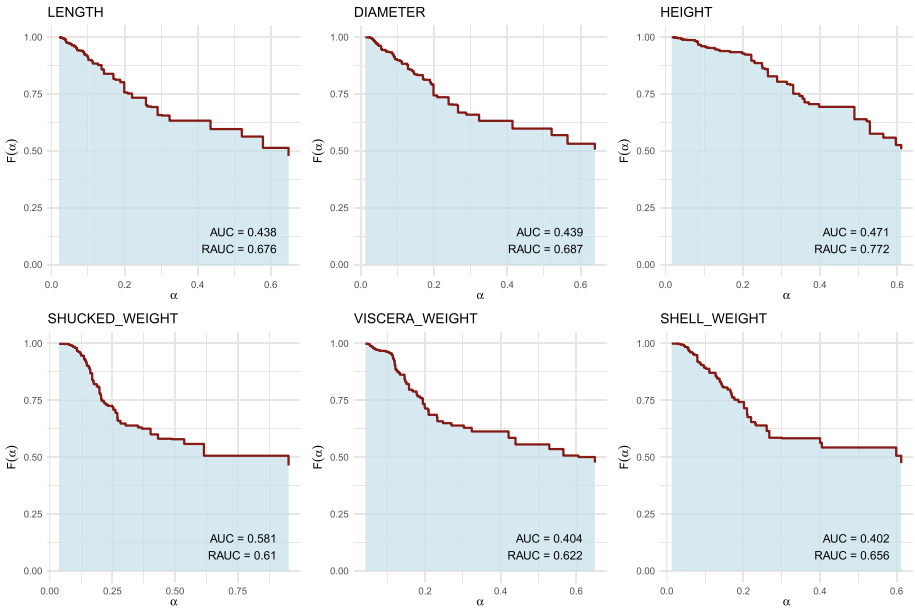


Fig. 3. Relative area under the curve (`AUCR`) of the average curve ϕ_j for each variable j of the `abalone` dataset.

RAUC values, meaning that the void space is reduced quickly, which corresponds to less persistent geometric structure.

4.2 Experiment 2

As in experiment 1, the dataset `Cardiologicalv2` of the `RSDA` package was used. This symbolic table has 44 observations of 5 interval-valued variables. The response variable is Pulse and we are using the other four variables as predictors: `Syst`, `Diast`, `Art1` and `Art2`.

In Fig. 4, the symbolic interval-based visualizations reveal distinct rectangular groupings for each predictor variable plotted against the response variable `Pulse`. These groups, which represent interval-valued data, exhibit varying structural characteristics; for example, the variable `Diast` shows one isolated (outlier) observation, while `Art2` has two outlier observations.

As shown in Fig. 5, the variables `Syst` and `Art1` exhibit relatively narrow bands across most of the α range, indicating a strong geometric alignment between the central trajectories and their associated dispersions. Notably, the $\phi - \psi$ and $\phi + \psi$ curves stabilize more quickly in these variables than in `Diast` and `Art2`, meaning that the process of filling the empty space occurs more slowly in the latter variables. Consequently, the geometric structure is more consistent in `Syst` and `Art1`.

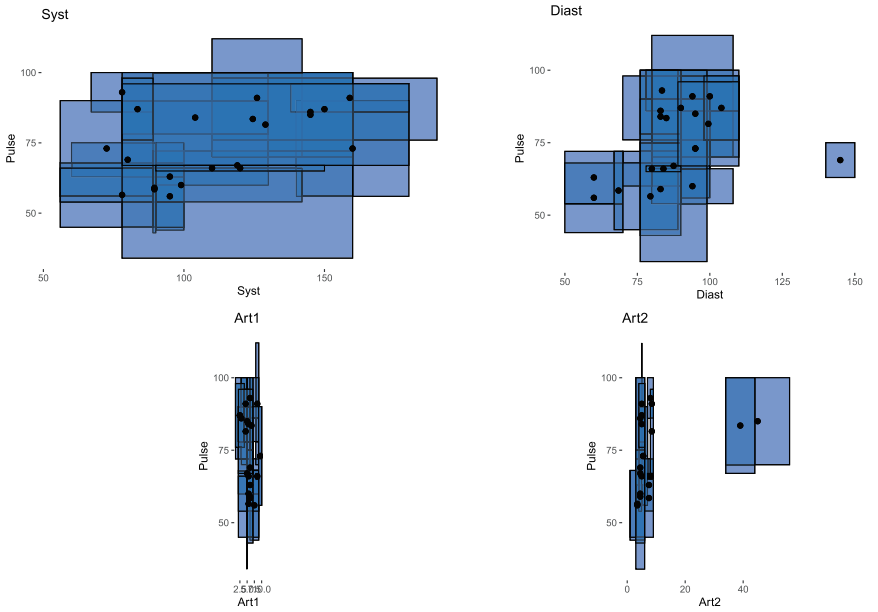


Fig. 4. Rectangular representations of interval ranges for each predictorâĂŞresponse pair in the interval dataset `Cardiologicalv2`, with a point cloud per pair showing the centers of the rectangles (N_j^c).

On the other hand, Fig. 3 presents the relative area under the curve, ϕ_j, for each variable j. It can be observed that `Syst` and `Art1` have the highest $RAUC$ values, indicating that their average curves ϕ_j remain elevated for a longer range before reaching stabilization. This behavior suggests that the void space is filled more gradually, reflecting stronger and more persistent geometric patterns. In contrast, variables such as `Diast` and `Art2` show lower $RAUC$ values, implying a faster reduction of the void space and, consequently, less stable geometric structures (Fig. 6).

A self-reproducible document containing all the codes used in this work, can be accessed at the following link: https://github.com/DylanBC09/goodness-of-fit.

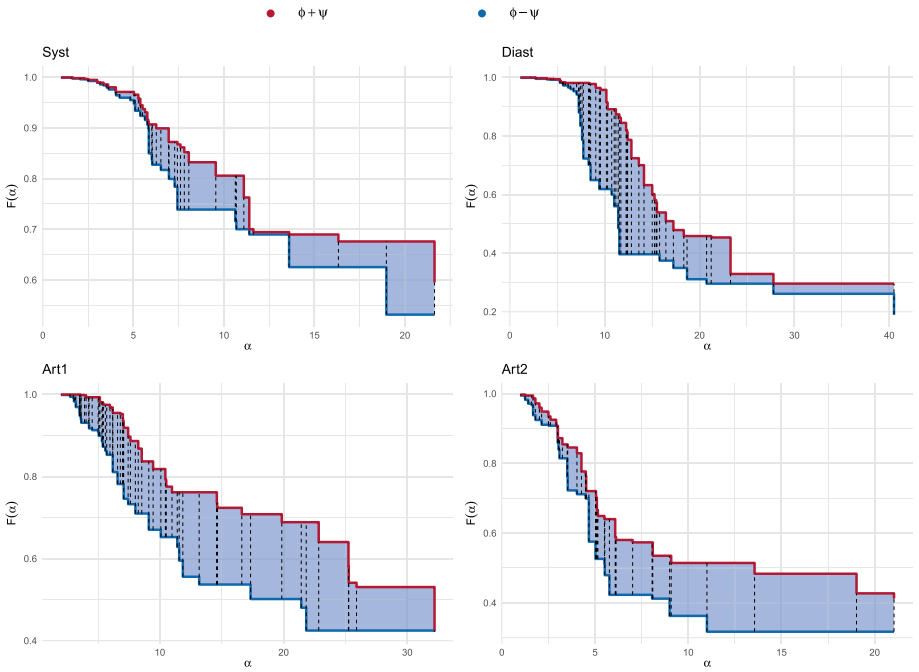


Fig. 5. Interval geometric fit function $F_j(\alpha)$ for each variable of the `Cardiologicalv2` dataset.

5 Limitations

The interval geometric fit function $F_j(\alpha)$ allows visualizing data structure at multiple scales, but its interpretation can become challenging when centers and ranges differ locally, as wide bands may coexist with high RAUC values. Moreover, F_j analyzes each predictor separately, without accounting for potential multivariate interactions. These limitations open opportunities for improvement, such as adaptively selecting α values or complementing F_j with metrics that more explicitly integrate the relationship between centers and ranges, enabling a more precise and robust interpretation in symbolic model evaluation.

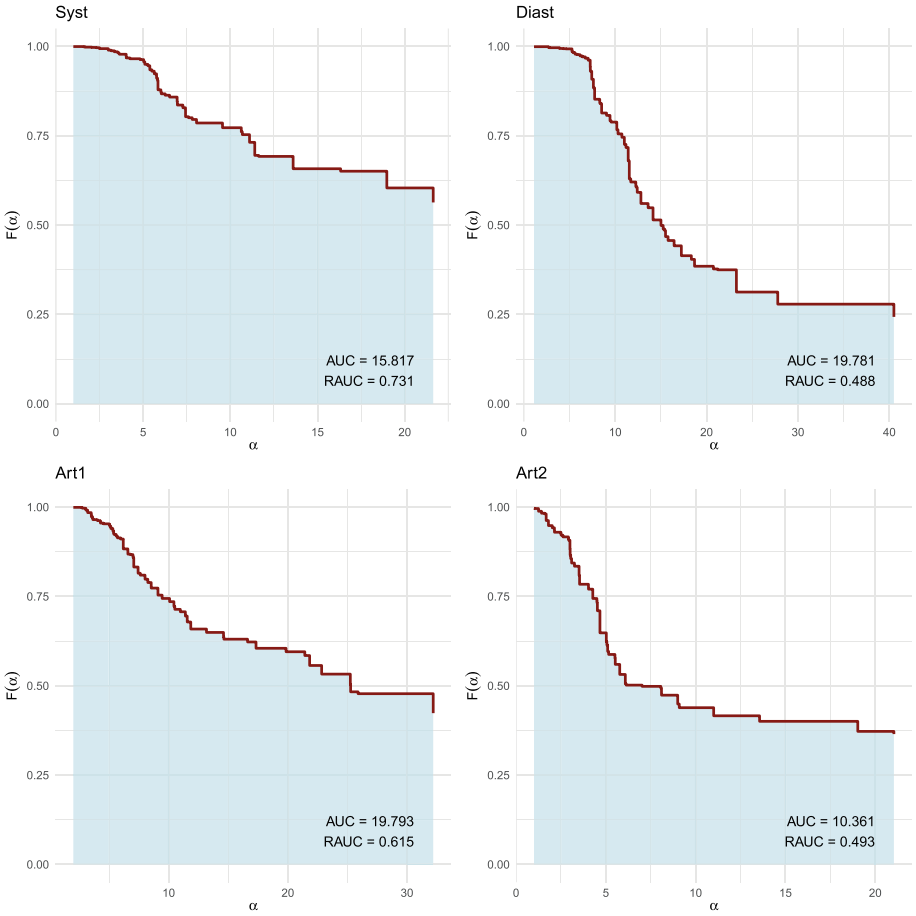


Fig. 6. Relative area under the curve (`AUCR`) of the average curve ϕ_j for each variable j of the `Cardiologicalv2` dataset.

6 Conclusions and Future Research

This study extended the geometric goodness-of-fit index originally proposed by Solís and Hernández (2023) to symbolic interval-valued data. Using alpha-shape constructions, we were able to capture the spatial structure of the rectangular clusters generated by each predictor and the response variable. The function F_j, defined in (4) provides multiscale insight into the presence or absence of geometric structure in the symbolic data. This function reflects how the spatial arrangement of rectangular clusters evolves as the parameter α increases, effectively generalizing the classical geometric goodness-of-fit curve to the interval-valued data.

The proposed approach effectively integrates both central tendency and variability of interval-valued data, highlighting non-randomness and identifying

potentially relevant predictors in models where traditional statistical indicators might fail to provide sufficient interpretability.

Future research will explore several directions. First, we aim to extend the geometric index to *histogram-valued* symbolic data, where each observation corresponds to a discrete distribution. This extension will require new methodological strategies for representing and analyzing histogram-based two-dimensional projections, potentially using sampling or optimal transport methods.

Second, integrating the geometric index into symbolic model selection and variable screening processes may enhance the interpretability and performance of symbolic regression models. Third, we propose to investigate the index's robustness and computational cost, developing data-driven procedures for choosing α values and exploring scalable implementations for high-dimensional datasets.

Finally, comparing this geometric measure with existing feature relevance techniques (e.g., Sobol indices) under symbolic settings will provide a broader perspective on its utility and limitations. The authors believe this geometric approach opens new avenues for analyzing symbolic data from a structural and interpretable standpoint.

Acknowledgments. This research was partially supported by a machine allocation on the Kabré supercomputer at the Costa Rica National High Technology Center (CeNAT), where the code for this study was developed. We would like to sincerely thank Maikol Solís and Oldemar Rodríguez for their valuable contributions throughout the research process.

Disclosure of Interests. The authors declare that they have no competing interests relevant to the content of this article.

References

1. Arce Garro, J., Rodríguez Rojas, O.: Optimized dimensionality reduction methods for interval-valued variables and their application to facial recognition. Entropy **21**(10), 1016 (2019). https://doi.org/10.3390/e21101016
2. Baddeley, A., Rubak, E., Turner, R.: Spatial Point Patterns: Methodology and Applications with R. CRC Press, Boca Raton (2015). https://doi.org/10.1201/b19708
3. Chacón, J.E., Rodríguez, O.: Regression models for symbolic interval-valued variables. Entropy **23**(4), 429 (2021). https://doi.org/10.3390/e23040429
4. Cramer, J.S.: Mean and variance of R^2 in small and moderate samples. J. Econometrics **35**(2–3), 253–266 (1987). https://doi.org/10.1016/0304-4076(87)90001-4
5. Edelsbrunner, H.: *A Short Course in Computational Geometry and Topology*. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-319-05957-0
6. Hernández, A.J., Solís, M.: Geometric goodness of fit measure to detect patterns in data point clouds. Comput. Stat. **38**(3), 1231–1253 (2023). https://doi.org/10.1007/s00180-022-01296-z
7. Lima Neto, E.A., De Carvalho, F.A.T.: Nonlinear regression applied to interval-valued data. Pattern Anal. Appl. **20**, 809–824 (2017). https://doi.org/10.1007/s10044-016-0561-6

8. Lima Neto, E.A., De Carvalho, F.A.T.: Constrained linear regression models for symbolic interval-valued variables. Comput. Stat. Data Anal. **54**(2), 333–347 (2010). https://doi.org/10.1016/j.csda.2009.08.018
9. Solís, M., Hernández, A.: UMAP projections and the survival of empty space: a geometric approach to high-dimensional data. In: Proceedings of the 18th International Federation of Classification Societies (2024)

Multi-objective Optimization for Strategic Shelter Placement: Case Study of Lima's Earthquake Preparedness

Soledad Espezúa(✉), Alexandra Sanjinez, Amy Checcllo, and Alexia Ríos

Universidad del Pacífico, Lima, Peru
s.espezual@up.edu.pe, {ac.sanjinezm,aa.checclloh,as.riosa}@alum.up.edu.pe

Abstract. This study addresses the optimization of temporary shelter locations for earthquake preparedness in Lima, Peru. A multi-objective approach is implemented using two evolutionary algorithms, NSGA-II and SPEA-II, to identify shelter sites that maximize population coverage, prioritize areas with high seismic vulnerability and risk, and ensure proximity to essential services such as healthcare facilities and water sources. The analysis integrates geospatial data on 5,855 potential shelters, evaluating scenarios under tuned algorithm configurations. Comparative results indicate that while SPEA-II achieves higher and more stable average fitness values, NSGA-II consistently reaches higher maximum objective values and exhibits greater diversity across the Pareto front, leading to a more balanced spatial distribution of shelters. For this reason, NSGA-II was selected for district-level and shelter characteristic analyses. The overlap with official municipal designations is limited (6 %), revealing differences in prioritization and highlighting underserved high-risk districts. An interactive web-based visualization was developed to support decision-making, enabling exploration of Pareto-optimal solutions and their spatial implications. The proposed framework offers a replicable methodology for seismic hazard–exposed urban areas, integrating optimization, spatial analysis, and decision support tools.

Keywords: earthquake resilience · NSGA-II · SPEA-II · temporary shelters

1 Introduction

Earthquakes, driven by the dynamic interactions of tectonic plates, are among the most destructive natural phenomena, causing significant loss of life and infrastructure damage [13]. Peru, located in the seismically active Pacific Ring of Fire, faces heightened risks, particularly in its coastal regions where tectonic activity is frequent [17,18]. The persistent threat of earthquakes underscores the need for robust disaster preparedness strategies to protect vulnerable populations.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 45–59, 2026.
https://doi.org/10.1007/978-3-032-20322-9_4

Lima, Peru's capital, is particularly at risk due to its location in a high-seismic zone and a prolonged period without major earthquakes, increasing the likelihood of a significant event in the near future [10]. Effective preparedness is critical, with temporary shelters playing a pivotal role in providing safe havens for displaced residents post-disaster [3]. These shelters must be strategically located to balance safety, accessibility, and resource availability, a challenge that requires advanced decision-making tools.

Optimizing the placement of temporary shelters involves addressing multiple, often conflicting, objectives, such as maximizing safety and ensuring access to critical resources. Recent studies have applied multi-objective optimization to enhance disaster response [4,19]. For example, Espejo [6] developed a model for shelter placement in Chile, focusing on risk scenarios, while Xu et al. [20] used Geographic Information Systems (GIS) to optimize shelter locations in Yangzhou, China, based on safety and travel time. However, these approaches often neglect key factors like access to healthcare and water, which are essential for effective post-disaster support.

Evolutionary algorithms are well-suited for this problem because they can explore trade-offs between multiple objectives and handle complex spatial constraints. The Non-dominated Sorting Genetic Algorithm II (NSGA-II) and the Strength Pareto Evolutionary Algorithm II (SPEA-II) are widely used in multi-objective optimization. For example, Aghaie [2] applied NSGA-II in Tehran to consider vulnerability and resource access, while Zhao et al. [21] minimized travel distances without integrating essential resources.

This study applies NSGA-II and SPEA-II to optimize the placement of temporary shelters in Lima, incorporating seismic vulnerability, population exposure, and access to essential services. The performance of both algorithms is compared to identify robust configurations. The approach integrates GIS to map vulnerability and population distribution, addressing gaps in previous studies by including resource accessibility as a key criterion alongside safety.

The structure of the paper is as follows: Sect. 2 introduces key concepts of multi-objective optimization and the NSGA-II and SPEA-II algorithms. Section 3 outlines the methodology. Section 4 presents the results and discussion. Finally, Sect. 5 concludes with key findings and implications.

2 Background

The Strength Pareto Evolutionary Algorithm 2 (SPEA-II) [22] is an advanced technique for multi-objective optimization, improving upon earlier methods by prioritizing a diverse set of non-dominated solutions. It uses a fitness evaluation that balances dominance and solution density, maintaining an archive of elite solutions to enhance diversity, which is valuable for complex decision-making scenarios.

The Non-Dominated Sorting Genetic Algorithm II (NSGA-II) [5] is a widely used approach that applies genetic algorithm principles to solve problems with multiple competing goals. It evolves a population of potential solutions through

selection, combination, and variation, using a ranking system based on dominance to identify the best trade-offs, known as the Pareto front. This process ensures a balanced set of options tailored to specific needs.

Multi-objective optimization [14] involves finding solutions that optimize several objectives simultaneously, often requiring trade-offs. Genetic algorithms (GAs) [8,9], the foundation of NSGA-II and SPEA-II, mimic natural evolution to explore these trade-offs effectively.

SPEA-II and NSGA-II complement each other in this study. NSGA-II excels at quickly identifying effective solutions, while SPEA-II enhances diversity, particularly in challenging optimization landscapes. Together, they provide a robust foundation for optimizing emergency shelter placement in Lima, addressing safety, population coverage, and resource access in a balanced manner.

3 Methodology

The optimization process for shelter placement in Lima consisted of four stages: data preparation, candidate shelter assessment, shelter localization using NSGA-II and SPEA-II, and result visualization (Fig. 1).

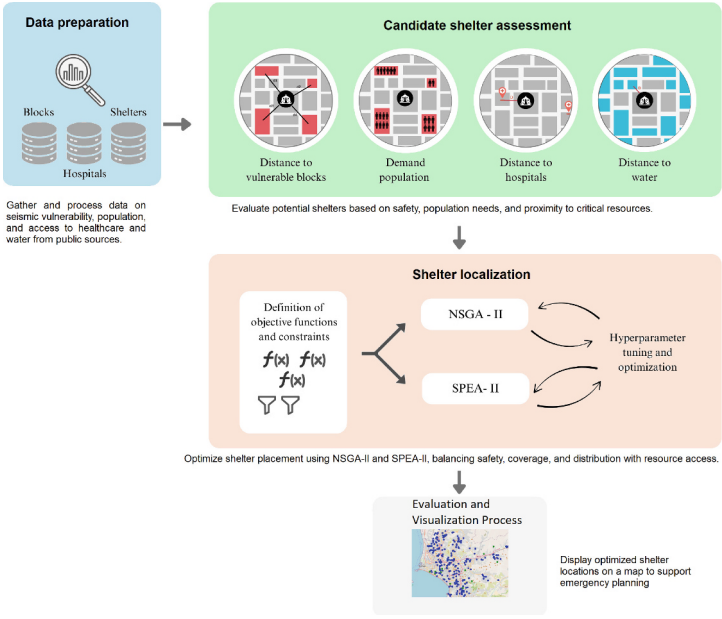


Fig. 1. Flowchart of the methodology.

3.1 Data Preparation

Data were collected from public repositories and geospatial sources [11,15,16]. The dataset includes seismic vulnerability, population distribution, and access to healthcare and water. Distances to high-vulnerability and high-risk blocks, healthcare facilities, and water sources were calculated using the Haversine formula. All distance variables were normalized using min-max scaling to allow integration in the optimization process.

3.2 Shelter Localization Using NSGA-II and SPEA-II Algorithms

The evolutionary algorithm was configured with a multi-objective fitness class oriented to maximization, assigning equal weights $(1.0, 1.0, 1.0)$ to the three objectives. Individuals, represented as lists, were initialized randomly to ensure diversity in the initial population. For genetic operators, Ordered Crossover (cxOrdered) [12] was applied to preserve the relative sequence of elements, while Shuffle Indexes mutation [7] introduced variability by randomly reordering genes with an independent probability of $p = 0.05$. This combination allowed both exploration of the search space and the prevention of premature convergence.

Each potential shelter was evaluated according to: (i) distance to high-risk zones, (ii) population demand within 1 km, and (iii) proximity to healthcare centers and water supply points.

Constraints included: a) minimum distance of 0.5 km from water-covered blocks and b) maximum distance of 1.5 km from healthcare facilities.

The main variables used in the assessment are listed in Table 1.

Table 1. Geospatial variables for shelter assessment

Variable	Description
Shelter ID	Unique identifier for each shelter
Shelter Name	Name of the shelter
District	District where the shelter is located
Latitude	Geographic latitude of the shelter
Longitude	Geographic longitude of the shelter
Area	Shelter area (m^2)
Blocks	Blocks within a 1 km radius
Normalized Vulnerability Distance	Sum of normalized distances to high-vulnerability blocks
Normalized Risk Distance	Sum of normalized distances to high-risk blocks
Demand Population	Population in high-risk/vulnerability blocks within 1 km
Hospital Distance	Distance to the nearest hospital (km)
Water Block Distance	Distance to the nearest water-covered block (km)

The optimization considered three objectives:

1. **Maximize inter-shelter distance** $f_1(\mathbf{x})$: Ensure distribution of shelters across the city for coverage.
2. **Maximize distance from high-risk areas** $f_2(\mathbf{x})$: Position shelters away from areas with high seismic vulnerability and risk.
3. **Maximize coverage of the population in need** $f_3(\mathbf{x})$: Prioritize shelters that can accommodate vulnerable individuals within a 1 km radius.

The objective functions and constraints are defined as follows:

– f_1**: Inter-Shelter Distance (distance)**

$$f_1(\mathbf{x}) = \max\left(\sum_{i<j} d_{ij}\right) \tag{1}$$

– f_2**: Seismic Vulnerability and Risk (vulnerability)**

$$\begin{aligned} f_2(\mathbf{x}) = {} & \max\left((1 - \text{dist_norm}_v) \times \left(1 - e^{-k \cdot \text{n_blocks}}\right)\right) \\ & + \max\left((1 - \text{dist_norm}_r) \times \left(1 - e^{-k \cdot \text{n_blocks}}\right)\right) \end{aligned} \tag{2}$$

– f_3**: Demanded Population (coverage)**

$$f_3(\mathbf{x}) = \max\left(\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij} \cdot p_i\right) \tag{3}$$

– *constraints*: **Hospital and water accessibility distances**

$$x \in \{\, x \mid \text{dist}_{hosp}(x) < 1.5 \ \wedge \ \text{dist}_{water}(x) < 0.5 \,\} \tag{4}$$

Both algorithms incorporated these objectives and constraints, and their parameters were tuned to improve efficiency and solution quality. Performance was compared across configurations, and the best solutions were integrated into a web-based visualization tool.

3.3 Evaluation and Visualization Process

The outputs from NSGA-II and SPEA-II were compared to assess trade-offs between safety, coverage, and accessibility. To facilitate interpretation, an interactive web platform [1] was developed, displaying Pareto-optimal solutions in a 3D scatter plot, where each axis corresponds to one objective: distance between shelters (f_1), seismic vulnerability and risk (f_2), and demanded population coverage (f_3).

As shown in Fig. 2, users can select a solution to visualize its spatial distribution on a map of Lima, with shelters marked as algorithm-selected (blue), matching both algorithm and Municipality (green), or Municipality-only (black). Key statistics such as average distances to hospitals and water sources, total vulnerable population covered, and number of districts served are also provided, supporting informed decision-making in emergency planning.

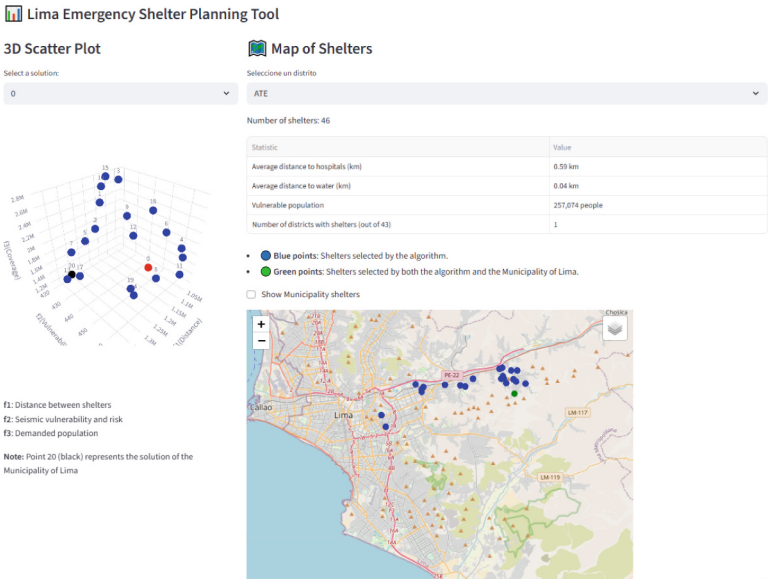


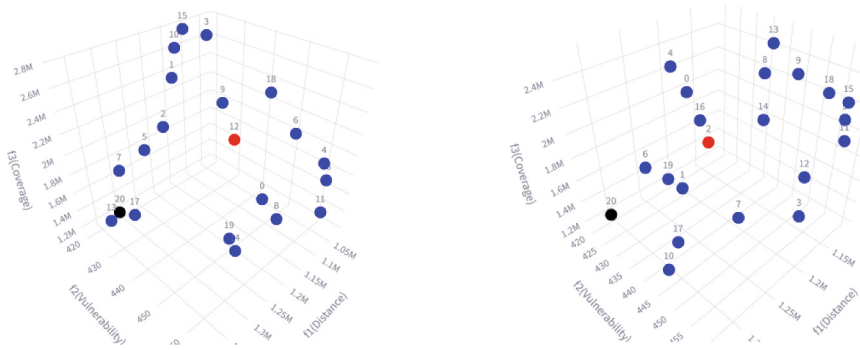
Fig. 2. Web interface for exploring Pareto-optimal shelter locations in Lima. Example in the map for district *Ate*: algorithm-selected shelters (blue) and matches with Municipality (green). (Color figure online)

4 Results

This section presents the results of the multi-objective optimization for shelter placement in Lima using NSGA-II and SPEA-II. It includes the parameter tuning process, the spatial distribution of optimized shelter locations, and a comparative analysis with official municipal designations. From an initial set of 5,835 potential shelters, 349 satisfied all spatial and service-accessibility constraints and were retained for the optimization process.

4.1 Parameter Tuning

Both algorithms were tested with different parameter combinations. Figure 3 shows the Pareto fronts for NSGA-II and SPEA-II. Blue points represent possible solutions, the red point is the selected solution, and the black point corresponds to the Municipality's selection. In both Pareto front visualizations, the municipal solution (black point) achieves high values for f_3 (coverage) but low values for f_1 (inter-shelter distance) and f_2 (distance from high-risk areas). This indicates a prioritization of population coverage at the expense of safety and spatial distribution. In contrast, the selected model solution (red point) achieves a more balanced trade-off, with intermediate-to-high values across the three objectives, avoiding extreme sacrifices in safety or spatial separation while maintaining adequate coverage.



(a) NSGA-II Pareto Front (b) SPEA-II Pareto Front

Fig. 3. Comparison of Pareto fronts for NSGA-II and SPEA-II. Blue points: non-dominated solutions; red point: selected model; black point: municipal solution. (Color figure online)

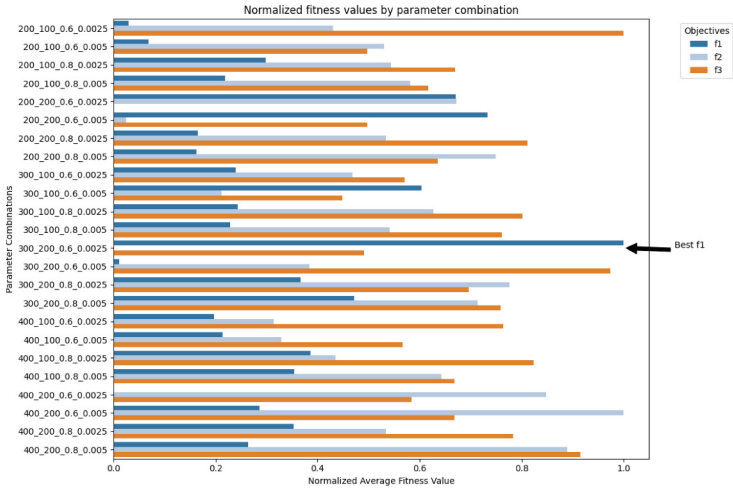


Fig. 4. Normalized average fitness values for different NSGA-II and SPEA-II parameter combinations. Note: The combinations follow the next order: population size, generations, crossover rate and mutation rate.

Figure 4 shows the normalized average fitness values for the 24 parameter combinations tested. In multi-objective optimization there is no single optimal solution; instead, the algorithms generate a set of non-dominated solutions, each representing trade-offs between f_1, f_2, and f_3. The figure reports average fitness over multiple runs.

Table 2. Normalized average fitness values for different NSGA-II and SPEA-II parameter combinations

Algorithm	Combination	$f_1(x)$	$f_2(x)$	$f_3(x)$
NSGA-II	200_100_0.6_0.0025	0.00	0.42	0.36
	200_100_0.6_0.0025	1.00	0.00	0.00
	200_100_0.6_0.005	0.26	0.98	0.13
	200_100_0.8_0.0025	**0.24**	**0.87**	**0.30**
SPEA-II	200_100_0.8_0.005	0.46	1.00	0.25
	200_200_0.6_0.0025	0.81	0.82	0.51
	200_200_0.6_0.005	0.73	0.72	0.82
	200_200_0.8_0.0025	**0.74**	**1.00**	**1.00**

Table 2 summarizes normalized average fitness values for each objective in the best-performing parameter sets. Higher values are highlighted in green. NSGA-II shows balanced performance, while SPEA-II performs better in specific trade-offs.

The best configuration for SPEA-II was a population size of 200, 200 generations, a crossover rate of 0.8, and a mutation rate of 0.0025, achieving the highest normalized values across the three objectives. This configuration was then applied to NSGA-II for comparison purposes in the subsequent analysis.

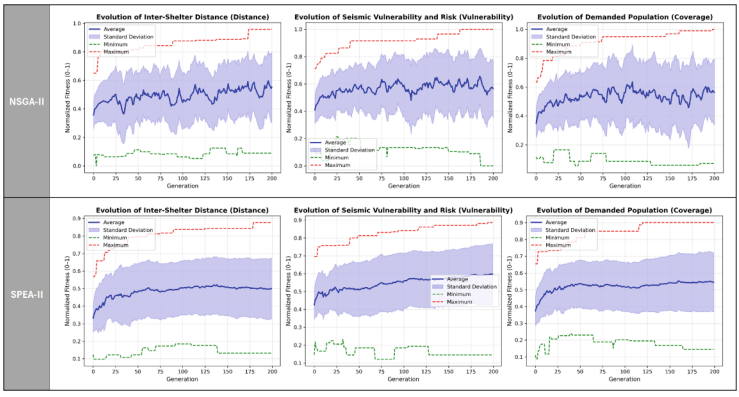


Fig. 5. Normalized average fitness values for different NSGA-II and SPEA-II parameter combinations.

Figure 5 compares the evolution of normalized fitness values for the three objectives—inter-shelter distance, seismic vulnerability and risk, and demanded population—under NSGA-II (top) and SPEA-II (bottom). Each plot shows the average fitness (solid blue line), variability (blue shaded area), and minimum and maximum values (green and red dashed lines). Both algorithms converge after approximately 100–120 generations, but their dynamics differ: SPEA-II

exhibits smoother and more stable average curves with lower dispersion, indicating efficient convergence, while NSGA-II shows higher variability yet consistently reaches greater maximum fitness values across all objectives.

This trade-off between stability and exploration is typical in multi-objective evolutionary algorithms: SPEA-II favors convergence and diversity preservation through its external archive and density-based fitness assignment, whereas NSGA-II promotes exploration by maintaining dominance-based ranking pressure and crowding distance sorting. Considering the application context—where spatial diversity and balanced trade-offs among objectives are critical for robust decision-making—the NSGA-II outcomes were selected for subsequent district-level and shelter-level analyses, as they provide wider Pareto coverage and a more evenly distributed spatial configuration of shelters across Lima.

4.2 Zone-Level Analysis

Lima was divided into four zones: North Lima, South Lima, East Lima, and Lima Center. Table 3 summarizes the number of blocks, percentage of blocks with very high vulnerability, percentage with very high risk, and number of districts in each zone.

Table 3. Description of the characteristics of the four zones in Lima

Zone	Number of blocks	% of blocks with very high vulnerability	% of blocks with very high risk	Number of districts
East Lima	37,358	26%	38%	9
North Lima	31,520	20%	28%	8
South Lima	25,313	21%	20%	12
Lima Center	12,779	9%	8%	14

Figure 6 compares the shelters selected by the Municipality of Lima and those obtained from the optimization models.

The Municipality's allocation resulted in 129 shelters in East Lima, 61 in North Lima, 66 in South Lima, and 93 in Lima Center. In contrast, NSGA-II assigned 109 shelters to East Lima, 97 to North Lima, 79 to South Lima, and 64 to Lima Center, while SPEA-II allocated 99, 100, 85, and 65 shelters to these zones, respectively. Both algorithms placed more shelters in high-vulnerability and high-risk areas—especially East Lima—while the Municipality concentrated more on central zones. Only 24 shelters overlapped between the Municipality's list and NSGA-II, and 18 with SPEA-II.

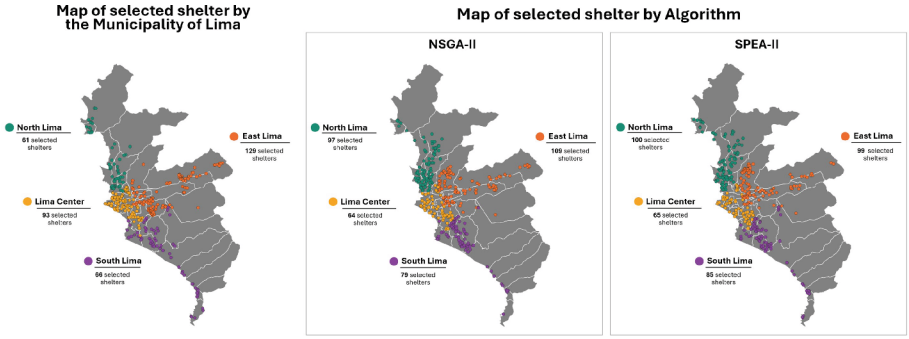


Fig. 6. Shelters selected by the Municipality of Lima (left) and by NSGA-II and SPEA-II (right).

4.3 District-Level Analysis

A district-level comparative analysis revealed notable differences between the shelters selected by the optimization algorithm and those proposed by the Municipality of Lima. The overall overlap between both selections was approximately 6%, meaning that only a small fraction of shelters were shared by the two approaches. In several districts—such as Ancón, Ate, Chaclacayo, Chorrillos, Cieneguilla, El Agustino, Independencia, Lurigancho, and San Juan de Lurigancho—there were no matches at all.

It is important to note that the Municipality's selection has certain strengths, particularly the placement of shelters in central areas with good road accessibility. However, some limitations were also identified, such as the absence of shelters in critical zones with high vulnerability and risk. In contrast, the optimization algorithm considered multiple criteria simultaneously, including population coverage, seismic vulnerability, the risk index, and the distance between shelters. This resulted in a spatial distribution better aligned with the needs of high-risk areas.

The limited overlap between the two methodologies highlights both differences in priorities and some points of convergence. Rather than replacing the municipal approach, the proposed model is intended to serve as a complementary tool, offering an evidence-based method to strengthen decision-making in disaster risk management.

Before conducting the district-level comparison, it is important to clarify the selection of the results presented. The parameter tuning phase showed that the best-performing configuration corresponded to SPEA-II, with a population size of 200, 200 generations, a crossover rate of 0.8, and a mutation rate of 0.0025. To ensure comparability, this configuration was also applied to NSGA-II. Although SPEA-II achieved higher average objective values, the Pareto front analysis revealed that NSGA-II offered a better spatial distribution of solutions, covering a broader trade-off space between objectives. For this reason, the district-level allocation analysis is presented using NSGA-II results (Table 4).

Table 4. Comparison of Municipality and NSGA-II results with matches

District	Municipality	NSGA-II	Matches (NSGA-II and Municipality)
Ancon	5	4	0
Ate	13	22	0
Barranco	2	1	0
Breña	6	2	0
Carabayllo	2	17	0
Chaclacayo	8	2	0
Chorrillos	9	9	1
Cieneguilla	3	1	0
Comas	6	14	0
El Agustino	3	9	0
Independencia	7	15	1
Jesús María	3	3	0
La Molina	55	5	2
La Victoria	4	2	0
Lima	23	6	1
Lince	4	3	1
Los Olivos	6	14	0
Lurigancho	33	18	6
Lurin	6	4	0
Magdalena del Mar	7	3	2
Miraflores	4	4	1
Pachacamac	6	1	1
Pucusana	5	1	0
Pueblo Libre	4	5	0
Puente Piedra	4	4	0
Punta Hermosa	3	2	0
Punta Negra	3	2	0
Rimac	11	6	1
San Bartolo	2	5	0
San Borja	4	5	0
San Isidro	4	3	1
San Juan de Lurigancho	5	43	1
San Juan de Miraflores	8	19	1
San Luis	3	4	0
San Martín de Porres	22	27	1
San Miguel	6	4	0
Santa Anita	6	5	1
Santa María del Mar	6	0	0
Santa Rosa	9	2	0
Santiago de Surco	10	13	0
Surquillo	3	5	1
Villa El Salvador	9	21	0
Villa María del Triunfo	7	14	1

An illustrative case is *Santa María del Mar*: although the Municipality lists six shelters, NSGA-II selects none. The district's small resident population (approximately 1,024 people) results in low demand within the $\sim$1 km catchment, causing candidate sites there to be dominated by alternatives that provide higher

coverage while also satisfying vulnerability, risk, and service-accessibility criteria. This outcome is consistent with the model's objective structure and may also reflect census-based population figures that do not account for seasonal fluctuations in occupancy.

4.4 Comparison of Shelter Characteristics

Table 5 compares the main attributes of selected and non-selected shelters. On average, selected shelters serve a larger demand population (7,691 vs. 4,679), present higher vulnerability (0.61 vs. 0.59) and risk (0.68 vs. 0.64) indices, and are located closer to both hospitals (0.45 km vs. 0.61 km) and water sources (0.054 km vs. 0.070 km).

Table 5. Comparison of selected and non-selected shelters based on various criteria.

	Selected	Non-Selected
Demanded Population	7691	4679
Vulnerability Index	0.61	0.59
Risk Index	0.68	0.64
Distance to Water Source	0.054	0.070
Distance to Hospital	0.45	0.61

These results suggest that the optimization process favored shelters in areas with higher vulnerability and seismic risk, ensuring proximity to critical resources. The shorter distances to hospitals and water sources indicate better accessibility to essential services during emergencies, while the higher demand population values reflect improved coverage of at-risk communities.

4.5 Computational Time

The computational time analysis was conducted exclusively for the NSGA-II algorithm, measuring execution times for each tested parameter configuration. Figure 7 presents the average times for different combinations of population size, number of generations, crossover probability, and mutation probability.

Population size emerged as the most influential factor, with configurations of 400 individuals showing the highest computational demands—often more than double those of smaller populations. The number of generations also had a direct and predictable impact, with 200 generations consistently requiring more time than 100 generations under the same conditions. These findings indicate that population processing and iterative generation stages are the primary contributors to the total computational cost.

In contrast, crossover and mutation probabilities had only marginal effects. Higher probabilities (0.8 crossover, 0.025 mutation) produced slight increases in

execution time, but their impact was negligible compared to population size and generation count. For time-sensitive applications, reducing these two parameters offers limited benefits; efficiency gains should instead focus on optimizing population and generation settings. However, smaller populations and fewer generations, while faster, may limit the algorithm's ability to explore the solution space thoroughly. A balanced configuration—such as a medium population size (around 300) with an appropriate number of generations—can provide a compromise between solution quality and execution time.

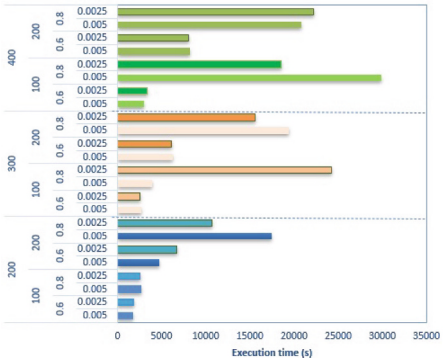


Fig. 7. Average execution times for NSGA-II under different parameter configurations.

In addition, it is important to note the computational cost of each algorithm. NSGA-II has a computational complexity of $\mathcal{O}(MN^2)$, where M is the number of objectives and N the population size, due to the non-dominated sorting procedure [5]. In contrast, SPEA-II reduces this cost to $\mathcal{O}(MN \log N)$ by incorporating an external archive and a more efficient fitness assignment, which makes it computationally less demanding for larger populations [22].

5 Conclusions

This study demonstrates that the multi-objective evolutionary algorithms can support the identification of strategic shelter location in Lima. Both NSGA-II and SPEA-II were evaluated with tuned parameters, with SPEA-II achieving the highest average objective value. However, analysis of the Pareto fronts revealed that NSGA-II provided a better spatial distribution of solutions, covering a broader trade-off space among the three objectives. For this reason, NSGA-II was selected for the district-level and shelter characteristic analyses presented in this work.

By integrating multiple criteria—population coverage, seismic vulnerability, seismic risk, proximity to health facilities, and access to water—the NSGA-II model generated a balanced distribution of shelters that effectively addresses

high-demand and high-risk areas. The district-level comparison revealed that, although the overlap with the Municipality's current selection is limited (6%) () 6%), the proposed approach complements existing planning by extending coverage to underserved critical zones.

The analysis of selected shelters showed that they tend to serve larger populations, are located in areas with slightly higher vulnerability and risk indices, and are closer to essential resources such as hospitals and water sources. These results indicate that the optimization process is aligned with operational priorities in disaster response.

Regarding computational time, the evaluation of different NSGA-II configurations highlighted population size and number of generations as the dominant factors influencing execution time. While higher values improve exploration of the solution space, they also significantly increase computation time, requiring a trade-off between efficiency and performance.

A limitation of this work is the reliance on 2017 Census data, which may not reflect current demographic or infrastructure changes. Future research could incorporate updated datasets, include dynamic variables such as real-time hazard forecasts, and extend the model to humanitarian logistics components—such as resource allocation, evacuation planning, and multi-hazard scenarios—enhancing its applicability for comprehensive disaster risk management.

Acknowledgments. We thank the Vice-Rectorate for Research of Universidad del Pacífico for the funding provided through the 4th Student Research Contest 2023, which made this project possible.

Data Availability Statement. The datasets and source code used in this study are available from the corresponding author upon reasonable request.

References

1. Web application for shelter visualization. https://simulator-7qpqcvtlhyalghkvym92m4.streamlit.app/. Accessed 08 Aug 2025
2. Aghaie, S., Karimi, B.: Location-allocation-routing for emergency shelters based on geographical information system (ArcGIS) by NSGA-II (case study: earthquake occurrence in Tehran (District-1)). Socio-Econ. Plann. Sci. **84**, 101420 (2022). https://doi.org/10.1016/j.seps.2022.101420
3. Castañeda, S., Vexlir, N.: Manual para la gestión y coordinación de albergues en el Perú (2021). https://peru.iom.int/es/resources/manual-para-la-gestion-y-coordinacion-de-albergues-en-el-peru
4. De Angeli, S., Malamud, B., Rossi, L., et al.: A multi-hazard framework for spatial-temporal impact analysis. Int. J. Disast. Risk Reduct. **73**, 102829 (2022). https://doi.org/10.1016/j.ijdrr.2022.102829
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
6. Espejo, L.G.A., Riesco, D.L.A.: Localización de albergues para hogares afectados por un desastre natural. In: Anais do XLII Simpósio Brasileiro de Pesquisa Operacional, Ubatuba (2011)

7. Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., Gagné, C.: DEAP: Distrib. Evol. algorithms Python—documentation. https://deap.readthedocs.io/en/master/api/tools.html. Accessed 27 Sep 2025
8. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc (1989)
9. Holland, J.: Adaptation in natural and artificial systems (1975)
10. Instituto Nacional de Defensa Civil: Escenario sísmico para Lima Metropolitana y Callao: Sismo 8.8 Mw (2017)
11. Instituto Nacional de Estadística e Informática: Población de la Provincia de Lima supera los 10 millones 292 mil habitantes (2024). https://m.inei.gob.pe/prensa/noticias/poblacion-de-la-provincia-de-lima-supera-los-10-millones-292-mil-habitantes-14869/. Accessed 08 Aug 2025
12. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. Multimedia Tools Appl. **80**(5), 8091–8126 (2021)
13. Lay, T., Kanamori, H.: Insights great 2011 Jpn Earthq. Phys. Today **65**(12), 33–39 (2012). Doi:https://doi.org/10.1063/PT.3.1817
14. Mirjalili, S., Dong, J.: Multi-Objective Optimization Using Artificial Intelligence Techniques. Springer, Cham (2020)
15. Municipalidad de Lima Metropolitana: Mapa de zonas seguras (2024). https://aplicativos.munlima.gob.pe/extranet/zonas/zonas-seguras
16. Municipalidad Metropolitana de Lima: Sistema de información metropolitana (2024). https://mmlimp.maps.arcgis.com/apps/webappviewer/index.html?id=ae3b2389664c4e199155e93a6aca9c25
17. Tavera, H.: Actualización del escenario por sismo, tsunami y exposición en la región central del Perú (2017)
18. Tavera, H.: Análisis y evaluación de los patrones de sismicidad y escenarios sísmicos en el borde occidental del Perú . Instituto Geofísico del Perú (2020)
19. Vollmer, M., Berchtold, C., Anniés, J.: Co-creating solutions for disaster risk reduction in multi-country research projects: opportunities and challenges. Int. J. Disast. Risk Reduct. **117**, 105187 (2025)
20. Xu, J., Yin, X., Chen, D., An, J., Nie, G.: Multi-criteria location model of earthquake evacuation shelters to aid in urban planning. Int. J. Disast. Risk Reduct. **20**, 51–62 (2016)
21. Zhao, M., Chen, Q., Ma, J., Cai, D.: Optimizing temporary rescue facility locations for large-scale urban environmental emergencies to improve public safety. J. Environ. Inf. **29**(1) (2017)
22. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength Pareto evolutionary algorithm. TIK Rep. **103** (2001)

Solving the Qubit Transmission Problem with Biologically Inspired Neural Networks and Reinforcement Learning

Ariel Fiuri[2(✉)], Martín Domínguez[2], Sofía Perón Santana[1,2], and Omar Osenda[1,2]

[1] Instituto de Física Enrique Gaviola (CONICET-UNC), Córdoba, Argentina
sofia.peron@mi.unc.edu.ar
[2] Facultad de Matemática, Astronomía, Física y Computación (UNC), Córdoba, Argentina
{ariel.fiuri,martin.dominguez,omar.osenda}@unc.edu.ar

Abstract. Over the last few years, machine learning methods have used deep neural network architectures to tackle complex problems. In this paper, we applied biologically inspired artificial neural machine learning to solve a specific task from quantum physics, i.e., quantum state transfer in a one-dimensional spin chain. The mission is to find transfer schemes with high transfer fidelities. We use different methods to learn a policy able to yields a good solution. As a first approach, we combine some architectures for Fiuri Artificial Neural Networks (FANN) with Reinforcement Learning (RL) and optimization techniques to train models, and as a second approach, we find solutions with different RL tabular methods. We use transfer fidelity as a metric to make a detailed comparison of the obtained results, combining different architectures and optimization methods for the FANN approach and different algorithms for the tabular one. We obtained very competitive results compared with the solutions provided in the literature.

Keywords: Artificial Neural Network · Reinforcement Learning · Optimization · Neural Dynamics · Biologically Inspired · Quantum Transmission State

1 Introduction

The outstanding control accuracy needed to implement quantum information processing in actual implementations is one of the practical obstacles to advancing in the application of quantum technologies to everyday problems.

Even in the absence of external perturbations, the controlled time evolution of quantum systems is not always achievable. For a closed quantum system, *i.e.*. without perturbations, the Hamiltonian operator dictates the time-evolution of the system, accordingly with the postulates of Quantum Mechanics. Quantum information processing involves three simple steps.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 60–78, 2026.
https://doi.org/10.1007/978-3-032-20322-9_5

First, prepare the initial state. Then, let the Hamiltonian drive the time evolution of the system until the quantum state contains the information with the answer to the problem of interest. The final step usually requires measuring the final state. In most cases, obtaining an accurate answer requires repeating the whole procedure. Sometimes, restricting the problem to study simplifies the task at hand. That is the case in the transmission (or transfer) of quantum states. In this context, an array of simple physical systems interacting with each other acts like a transmission channel. The preferred geometry for this channel is a linear array or chain. A chain of qubits is an array of the simplest quantum mechanical systems. From the mathematical point of view, there is no simpler system than one whose quantum state belongs to a two-dimensional complex vector space, as is the case with a qubit.

So, the system of interest is a chain of equally spaced qubits. The chain has a Hamiltonian independent of time. Besides, we consider that a given qubit interacts only with its two closest neighbors. For this setting, the only way to successfully transfer a quantum state from one extreme of the chain to the other is by controlling external actions applied over the qubits, except for a handful of cases where perfect state transfer results from the time evolution given by the time-independent Hamiltonian. The simplest protocol of transmission only considers the transmission of an arbitrary one-qubit state. In this case, the quantum state of the qubit chain is a vector with N complex components, and the Hamiltonian is an NxN Hermitian matrix, for a chain with N qubits. So, to control the time evolution of the chain, we should consider the number and type of actions applied. Guided by the actions implementable in actual physical systems, we will consider a reduced number of actions.

There are numerous studies about controlled quantum state transfer using different methods to design the controlling actions. Older studies employ different implementations of Optimal Control Theory (OCT), based on the Krotov method. More recently, many studies exploit the flexibility of Machine Learning methods to produce promising control sequences. OCT is a well-understood framework, with a pleasing physical interpretation. Regrettably, OCT is very demanding in computing time when the number of actuators is large.

Machine learning methods, such as the genetic algorithm or Reinforcement Learning, possess better scalability when dealing with a considerable number of control actions. In particular, deep reinforcement learning applied to the design of control policies in distinct problems has shown great promise. Nevertheless, some issues remain concerning, and further investigation is needed. The architecture of the underlying neural networks, training policy, and usefulness of the trained model are some of the issues that need clarification.

Biologically inspired artificial neural networks have revolutionized the field of machine learning and neuroscience, providing models capable of learning and adapting efficiently. Inspired by the functioning of the animal nervous systems behavior, these networks mimic neuronal connections and processes to tackle complex problems. In the context of quantum physics, where intrinsic nonlinearity and high dimensionality pose significant challenges, biologically inspired

artificial neural networks emerge as promising tools for function approximation, model optimization, and predicting quantum behaviors. In this frame, a limited version of the FANN were successfully used to solve RL related learning situations [3]. We based on that results evolve the technique to expand to different architectures and also to solve a novel problem for this metodology. So we extend that idea using them to solve quantum state transfer in a one-dimensional spin chain problem. Our study focused on two key aspects: the architecture of the neural network and the learning mechanism. Our main objective was to address the problem of understanding the relative capabilities of biological inspiration as far as using it to efficiently solve specific problems unrelated to RL. Throughout this process, we successfully tackled the problem based on accumulated experience using a small network. Our main contribution is the extension of the Fiuri neural dynamics model and FANNs (biologically inspired networks), pushing their applicability both in terms of architectural design and the complexity of problems they can address. Additionally, we provide a significant advancement in the quantum control domain, particularly regarding knowledge representation. We introduce two novel alternatives to existing approaches: leveraging the real part of quantum states and adopting a condensed information scheme to feed agent inputs. These innovations drastically reduced computation times while preserving competitive performance. Altogether, our results strengthen the position of FANNs as a promising and competitive alternative to state-of-the-art methods.

2 Background

2.1 Qbits

Here, we present the control scheme used to achieve high-quality quantum state transfer, the mathematical expression of the Hamiltonian, the evolution operator, and the necessary details to formulate the control problem as a learning problem.

The time dependent quantum state, $|\psi(t)\rangle$, obeys the Schrödinger Equation

$$i\frac{d|\psi(t)\rangle}{dt} = H|\psi(t)\rangle, \tag{1}$$

where H is the Hamiltonian of the system. For a time-independent Hamiltonian, the solution of Eq. (1) is trivial, $|\psi(t)\rangle = \exp(iHt)|\psi(0)\rangle = U(t)|\psi(0)\rangle$, where U is the evolution operator, $|\psi(0)\rangle$ is the initial quantum state, and the exponential depends on the spectral decomposition of H. Even for time-dependent Hamiltonians, the solution of Eq. (1) is given by $|\psi(t)\rangle = \tilde{U}(t)|\psi(0)\rangle$, but the evolution operator has a more convoluted expression than the simple exponential. It is important to note that since the Hamiltonian is a Hermitian operator, the evolution operator is unitary, so the time evolution preserves the normalization of the vector state $|\psi(t)\rangle$, which is consistent with the usual probabilistic interpretation associated with it.

When the parameters that define the Hamiltonian depend on time, but in a piecewise fashion, the quantum state at time t is given by

$$\begin{aligned} |\psi(t)\rangle &= \exp\left(iH_n\tau_n\right)\exp\left(iH_{n-1}\tau_{n-1}\right)\ldots\exp\left(iH_2\tau_2\right)\exp\left(iH_1\tau_1\right)|\psi(0)\rangle \\ &= U_nU_{n-1}\ldots U_2U_1|\psi(0\rangle = U|\psi(0)\rangle, \end{aligned} \tag{2}$$

where H_k is the appropriate time-independent Hamiltonian valid on the interval τ_k and $t = \sum_{k=1}^{n}\tau_k$.

As is customary in quantum mechanics, the Hamiltonian of a system of qubits is expressed in terms of the Pauli matrices since they form, together with the identity matrix, a basis set for all the Hermitian matrices operating on $\mathbb{C}^2$. So, a rather general Hamiltonian for a chain of N qubits valid at the τ_j interval is given by

$$H_j = \sum_{i=1}^{N-1}\left(\sigma_i^x\sigma_{i+1}^x + \sigma_i^y\sigma_{i+1}^y\right) + \sum_{k=1}^{N}B_{k,j}\sigma_k^z = H_{XX} + \sum_{k=1}^{N}B_{k,j}\sigma_k^z, \tag{3}$$

where $B_{k,j}$ is the control applied over the k-th qubit, σ_i^x, σ_i^y, and σ_i^z, are the Pauli matrices associated to the i-th qubit. In many-body physics, the Hamiltonian in Eq. (3) is called the XX Hamiltonian with site-dependent transversal magnetic field. We do not assume a specific physical setting associated with the Hamiltonian, but we assume that the system has tunable Z one-qubit gates, using the common terminology in quantum computation. The Hamiltonian in Eq. (3) operates in the space $\mathbb{C}^{2^N} = \bigotimes_{k=1}^{N}\mathbb{C}^2$. Nevertheless, since we aim to study the transmission of arbitrary one-qubit quantum states, we use the Hamiltonian matrix that results from reducing the Hamiltonian to the one-excitation subspace, spanned by the computational basis $|\mathbf{i}\rangle = |00\ldots1\ldots\rangle$, for $i = 1, 2, \ldots, N$, where the ket $|\mathbf{i}\rangle$ denotes a vector with N components, with only one component different from zero, equal to the unity located in the i-th component. Using this notation, it is straightforward to show that the Hamiltonian matrix on the one-excitation subspace is a tridiagonal one. The elements on the diagonal are the magnetic fields. All the other non-null elements are equal to unity.

We will use as the **actions** driving the evolution of the qubit chain the control "fields" $B_{l,k}$, with the proviso that for each time interval only one of the fields will be non-zero.

The average fidelity of transmission quantifies the quality of the transfer protocol and is given by $f_{av} = \frac{1}{2} + \frac{1}{3}|P| + \frac{1}{6}|P|^2$, where the probability of transmission is

$$|P|^2 = |\langle\mathbf{N}|U(t)|\mathbf{1}\rangle|^2. \tag{4}$$

Numerous methods allow for selecting sequences of fields $B_{j,k}$ which result in values of P close to unity. We aim to produce sequences compatible with fast transmission. For the units chosen for the Hamiltonian in Eq. 3, the sequences and methods employed produce arrival times T such that $N/2 \leq T \leq N$. For the arrival time, we understand the time that is smaller than N at which the transmission probability attains its maximum value. The number of time intervals τ_k is a hyperparameter of the particular method used.

2.2 Reinforcement Learning

Reinforcement Learning (**RL**) is an automatic learning technique in which an agent learns to perform a task through a succession of trial-and-error interactions with a changing environment, usually modeled as a Markov decision process. To train the agent in solving a problem, the environment provides us with a reward and a detailed description of the problem. Every time we execute an action on the environment, we obtain a reward and a descriptor of the State again. A desired solution maximizes the rewards obtained in a given number of interactions [15]. An episode refers to the sequence of interactions with the environment that begins from its initial state. In each episode, the rewards obtained in each interaction accumulate to calculate a total reward for that episode.

To solve the problem of **QBits** using **RL**, it was necessary to model and implement an environment responsible for providing some type of state description (**state**), a set of actions that can be applied to it ($\mathcal{A}$), a reward corresponding to the action executed (**r**) and a sign of completion (to know when the episode concludes). We define four different reward functions with the aim of exploring better solutions.

In this work, we address the problem of **QBits** through two **RL** general approaches: the first consists of tabular methods, and the second uses an extended **FANN** methodology presented in [3].

2.3 Tabular RL

Q-learning is one of the most popular reinforcement learning algorithms used in automatic learning and artificial intelligence. Allows an agent to learn the optimal actions that must be taken in a given environment to maximize accumulated rewards over time. A key concept in **Q-Learning** is that of the values Q (Value-action function), responsible for storing estimates of the expected accumulated reward that the agent can obtain when executing a specific action in a given state. These values are then used to generate a policy of selection of actions, as presented in [15]: $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$ where: s and s' are the current and next states, a and a' are actions, r is the received reward, α is the learning rate and γ is the discount factor.

The agent chooses actions based on an exploration strategy, such as *Epsilon-Greedy*, to balance the exploration of new actions and the exploitation of actions already known as good. Over time, the values converge towards optimal value-action, guiding the agent to make optimal decisions. In this work, we refer to this model as **QN**.

Double Q-Learning is an extension of the standard **Q-Learning** algorithm, designed to address overestimation bias that can occur in traditional **Q-Learning**. This bias occurs because the maximum operator in the **Q-Learning** update tends to overestimate the value of the actions, especially when the values are low or approximate. **Double Q-learning** maintains two separate estimates of Q values instead of one. These are commonly denoted as Q_1 and Q_2.

The update of the Q values, after executing one action and observing the reward and the following state at each step of the episode, is carried out by randomly selecting one of the two functions Q (for example, q_1 or q_2) and using the other to determine the optimal action in the following state. The update rule for the selected Q function is as follows: $Q_i(s,a) \leftarrow Q_i(s,a) + \alpha\left[r + \gamma Q_j\left(s', \arg\max_{a'} Q_i(s',a')\right) - Q_i(s,a)\right]$ where: i and j are 1 or 2, with $i \neq j$. The idea is that Q_i is updated using the estimate from Q_j for the next state's maximum action, reducing overestimation bias. Like **QN**, the Q-values converge to the optimal action-values, guiding the agent to make optimal decisions. In this work, we refer to this model as **DoubleQN**.

Monte Carlo is the last tabular method used in this work (called us as **MC**). It is a direct approach to reinforcement learning, used to estimate the value of states or the state-action pairs from the average of complete episodes. It is especially useful in environments with spaces of finite and discrete actions, where a table can store these estimates. The agent follows a policy to generate episodes of experience. Then, for each state visited (or status-action), the return G_T is calculated as the sum of discounted rewards from that moment until the episode ends: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$

The value estimate for each state (or state-action pair) is updated by averaging all observed returns: $V(s) \leftarrow$ average of all G_t where s was visited.

2.4 FANN

Fiuri Artificial Neural Networks (**FANN**) are biologically inspired artificial neural networks as presented in [3], but in this work, we extended that method to some kind of fully connected architecture, which allows us to have multiple intermediate layers. We added a new connection behavior, trying to simulate the modulation mechanism in the neuron information transmission (that version of the model includes an **M** in its name). A **FANN** is a compound of neurons with a discrete dynamic, which is in charge of describing its behavior. The interneuronal connections also have a well-defined behavior depending on their type (excitatory, inhibitory, or electrical). That dynamics (shown in Table 1) describes how the neuron reacts when stimuli are coming from its dendritic connections. We use the well-studied **Tap Withdrawal Circuit** (**TWC**) architecture that models a small portion of the *C-elegance* connectome [4,16]. This small network consists of eleven neurons, with four sensory, two motor, and five internal neurons. We also use a variant of that circuit named **Shuffle Architecture** (**SA**) obtained from the same neurons as **TWC** but with all its connections randomly shuffled. The Table 2 shows a graphical view for both of them.

Then we use the **TWC** as [3] in its two versions (including Shuffle Architecture **SATWC**) but also our extension to fully connected architecture named **FC** (and **SAFC** for the shuffle architecture version). We can think of **FC** as layers where we can stimulate the neurons belonging to the input layer. Each neuron of

Table 1. Neural Dynamics. In ***Fiuri***, E_n and O_n represent the input and output states of neuron n, respectively, once it was influenced. If S_n represents the stimulus coming through the dendritic connections ($j's$) due to the currents I^j_{in} (7), then the expressions (5) and (6) define the dynamics of n. T_n and d_n are neuronal parameters that have to be learned and represent the firing threshold and the decay factor (due to not enough stimulus). The currents I^i_{in} are defined by (8) for each connection type. ω_j is the connection capacitance.

$$O_n = \begin{cases} S_n - T_n & \text{if } S_n > T_n \\ 0 & \text{other case} \end{cases} \quad (5)$$

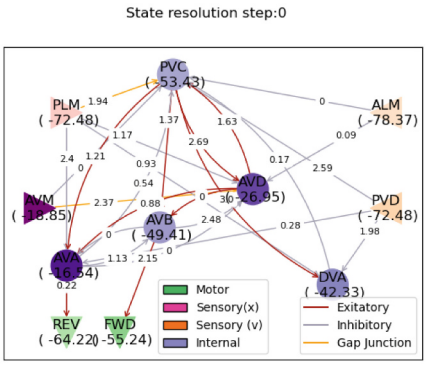
$$S_n = E_n + \sum_{j=1}^{m} I^j_{in} \quad (7)$$

$$E_n = \begin{cases} S_n - T_n & \text{if } S_n > T_n \\ E_n - d_n & \text{if } S_n \leq T_n \text{ and } S_n = E_n \\ S_n & \text{other case} \end{cases} \quad (6)$$

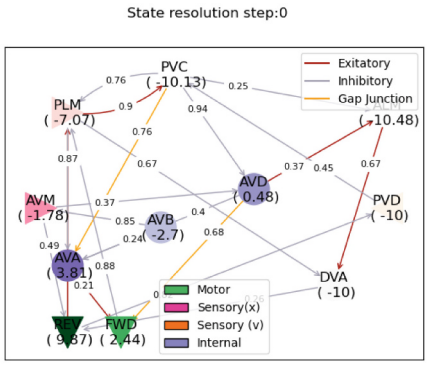
$$I^j_{in} = \begin{cases} \omega_j * O_j & \text{if } O_j \geq E_n \text{ y gap junct.} \\ -\omega_j * O_j & \text{if } O_j < E_n \text{ y gap junct.} \\ \omega_j * O_j & \text{chemical excitatory} \\ -\omega_j * O_j & \text{chemical inhibitory} \end{cases} \quad (8)$$

this layer is connected to all neurons belonging to the first internal layer. Each internal layer has a neuron connected to the next internal layer, and so on till the last internal layer is connected in the same way to the output layer. The left image in Table 3 shows this; notice is similar to deep learning fully connected networks.

Table 2. Architecture examples: Numbers on neurons are the electrical voltage at the moment the picture was taken, and numbers on connections are their capacitance.

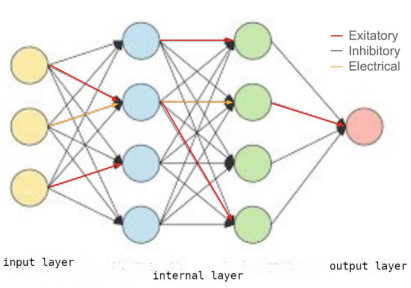


Tap Withdrawal Circuit Architecture.

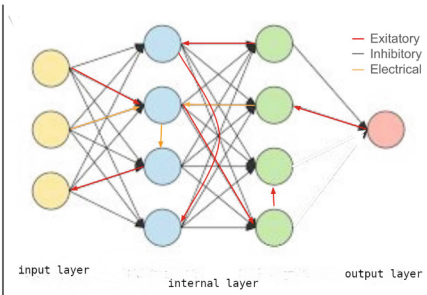


Example of **Shuffle Architecture**.

Table 3. Architecture examples. Notice SAFC allows loops.

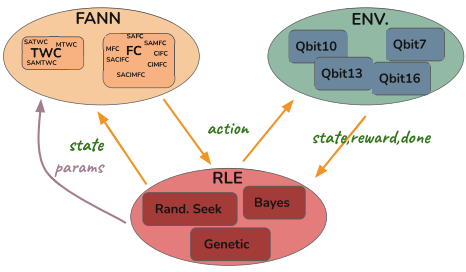


FC Architecture.



Example of **SAFC**.

SA can break this order due to the connections being randomly shuffled, even allowing loops as shown in the right figure. That kind of network maintains the bio inspiration due to use the Fiuri neuron dynamic model, different connection types, the proportion of connections for each type (inhibitory, excitatory, and electrical), and the new modulation mechanism. The Fig. 1 attempts to help with a schematic view.



All these architectures learn due to the optimization procedure on the ANN internal parameters, so we also maintain the three mechanisms used in [3] (Random Seek, Genetic, and Bayesian) RL in this method simply acts as a "circuit evaluator" due to only accumulating the obtained step rewards on each episode, storing the ANN that has the best total reward.

Fig. 1. Schematic view.

3 Related Works

References [2,9,10] studied controlled quantum state transfer using optimal control theory, although in different settings. Later studies found that quantum state transfer occurs at times close to the *quantum speed limit*, but early studies found arrival times orders of magnitude larger. Choosing a simple set of actions, Zhang et al. [17] achieved fast transfer times, with a good to poor transmission probability in qubit chains with short to moderate lengths. These authors used Deep Reinforcement Learning with an ϵ-greedy policy. The transmission of quantum

states through a RL designed chain with site-dependent interactions without external controls is the subject of reference [14]. Preliminary results [11,13] convinced us that exploring other RL strategies or different neural network settings was worthwhile. Improving the results of Zhang et al. using the same approach is indeed quite laborious. A closely related study is presented in [12], where the authors use ensemble reinforcement learning, while in [7] the authors present a variational quantum optimal control method. In [3], the methods employed in the present work were presented and developed thoroughly.

4 Our Approach

4.1 Exploring the Solution Space on FANNs

As mentioned before, traversing the space of potential solutions (trying different values for internal dynamic parameters) is crucial when training our **FANN** model. In this paper, we employ three well-known methods for exploring the parameter search space: **Random Seek** (**RS**), **Genetic Algorithm** (**GA**) [8], and **Bayesian Optimization** (**BO**) [5,6]. In all cases, we maintained the same hyperparameter optimization as in [3] remained fixed throughout the experiments.

4.2 Environment

A critical piece of any RL system is the environment that is in charge of modeling the problem we want to solve. We implemented this piece of code following the problem characteristics introduced before. We considered four versions of the same problem, named **Qbit7**, **Qbit10**, **Qbit13**, and **Qbit16**, which correspond to chain lengths of 7, 10, 13, and 16, respectively. Our environment has as many actions as the chain length (i.e. $\mathcal{A}$), so when we want to solve **Qbit7**, we will need to take into account that we have 7 actions, but if we want to solve **Qbit13**, we then have 13 actions. We assigned numbers $1..n$ to the actions and implemented them in such a way that the effect of executing the action n on the environment is to influence the chain link n.

Another key component is the *reward* the environment provides on each action execution. We define an ideal goal fidelity to be reached on every problem, for **Qbit7** is 0.99, for **Qbit10** is 0.98 for **Qbit13** is 0.95, and for **Qbit16** 0.92. Notice that when the chain length increases, fidelity degradation is expected. We proposed four different reward functions: **site evolution**, **transmission evolution**, **last action auto-complete** and **random action auto-complete**, here is a detail:

Site evolution(**SE**): aims to measure the position of the excited (up) state along the chain. To do that, we define the mean site as: $\mu = \sum |\alpha_i(t)|^2 i$ where $\alpha_i(t)$ are the coefficients of the system's full quantum state $|\psi(t)\rangle = \sum_i \alpha_i(t)\,|i\rangle$. Since transmission probability does not start to grow until several time steps have gone by, this reward tries to encourage actions that make the excited state

move forward along the chain, even in earlier stages where it is still far from the final qubit. The formula also includes the transmission probability, resulting in: $r(t) = (1 + P)\mu$.

Transmission evolution(**TE**): tries to induce the information wave moving to the last chain link. During each action execution, it maintains the current wave position (i.e., the specific chain link) and its amplitude. It rewards when the wave picks change one chain link forward to the final, or if no position change occurs, then it at least increases its amplitude. It penalizes in other cases. It returns on each step 10 if it was a good action, 50000 if it exceeded the objective, and 0.1 in other cases.

Last action auto-complete(**LA**): calculates the whole episode action list fidelity, but starting from the current state and auto-completing till the final with the last executed action. To be more clear, suppose that we are trying to solve **Qbit7**, so the episode is done when we execute 35 actions. In that example, suppose we want to execute action n and we have already run m actions on the environment, which is in the state e. This reward function will complete the episode (from state e), executing always n till reaches the *done* signal ($35 - m$ times). This reward function returns the value defined by (9) where o_n is the n-sized chain objective and fid the fidelity.

Random action auto-complete(**RA**): is like last action auto-complete, but in every step, the action used to complete the action list is random. This reward function also returns the value defined by (9) where o_n is the n-sized chain objective and fid the fidelity.

$$rew = \begin{cases} fid & \text{if } fid \leq 0.8 \\ 100/(1 + e^{(10*(o_{chain} - fid))}) & \text{if } 0.8 \leq fid \leq o_n \\ 50000 + fid & \text{if } o_n \leq fid \end{cases} \tag{9}$$

Another important piece is the state description that the environment provides us. We use as a state descriptor a complex n_uple representing the whole quantum chain state for a chain of size n. By last, we use a *done* signal equivalent to 5 times the chain length, which seems to be good enough.

4.3 Interfaces

The interface in that work concerned with the way the environment state is passed to the learning mechanism and how the learning mechanism's output is interpreted as an action. Our first approach for **FC FANN**s was to feed two input neurons with every chain link state; we passed the real part to the first and the imaginary part to the second one. So to train a **FC** trying to solve a chain with length n, we needed $2n$ input neurons. It was quickly noticed that passing both components to **FANN** was not necessary, and so we tried to pass only the real part of the state (in the same way, we attempted to pass only the imaginary part, with similar results). Therefore, for every chain link, we only have one input

neuron. In this way, the results were very similar to the first approach, but the train times were reduced drastically due to we needed n input neurons for the **FC**. A third approach we also tried was to have only 2 input neurons and feed the first with the wave peak amplitude and the second with the peak position (in terms of the chain link). A model with this kind of interface has a **CI** (Condensed Information) word in its name. This last approach reduced the training times once again due to only having a few internal parameters in comparison with others. For **TWC**, we can not select the number of input neurons, so we took the **CI** approach, and the same for tabular methods. Concerning the output neurons, all our **FC** networks have only one output neuron, which is directly matched with an action after a discretization step. In **TWC** nets, we have 2 output neurons that have to compete, and the result is discretized and mapped to an action.

4.4 Modulation Mechanism on FANNs

We added a modulation mechanism to the connection dynamics, just with the utterance of mimicking the biological modulation behavior in dendritic connections. Without that mechanism, the dendritic connection influence was defined by (8), so we present now how the neuron dynamics are influenced. The expression (10) shows that the neuron suffered influence when we use the modulation mechanism. Notice a new parameter σ is introduced there, this parameter needs to be learned so models with that mechanism (i.e., with **M** appearing in the name) have more parameters than others. This modulation is based on a sigmoid function and does not affect the electrical connections, which is also a biological inspiration [1].

$$I_{in}^{j} = \begin{cases} \omega_j * O_j & \text{if } O_j \geq E_n \text{ y gap junct.} \\ -\omega_j * O_j & \text{if } O_j < E_n \text{ y gap junct.} \\ \omega_j * O_j / (1 + e^{-\sigma * O_j}) & \text{chemical excitatory} \\ -\omega_j * O_j / (1 + e^{-\sigma * O_j}) & \text{chemical inhibitory} \end{cases} \tag{10}$$

4.5 FANN's Architecture Notes

We need to mention some considerations here. In the first place, we tried to maintain the rate between input and other neurons between 0.57 and 0.7, the first because it is observed in **TWC** and the second risen from experience. In a second place and the same sense, we tried to maintain the rate of connections by type concerning **TWC**, defined 55% of connections should be inhibitory, 35% should be excitatory, and approximately. 10% electrical. All these rates are rounded as needed, so they are not exact numbers. With all those considerations presented till moment, the Table 4 resumes the different architectures we explored, together with their names, and a brief description. Notice that the size architecture depends on the chain length in some configurations. Table 5 shows detailed info about configuration characteristics and their dynamical parameters.

Table 4. Different architectures for the explored FANNs

Architecture	Description
TWC	Clasical TWC
$SATWC$	Shuffle Architecture of TWC
$MTWC$	Clasical TWC with Modulation
$MSATWC$	Shuffle Architecture of TWC with Modulation
FC	Fully Conected FANN
MFC	Fully Conected FANN with Modulation
$CIFC$	Fully Conected FANN with Condensed Information
$CIMFC$	Fully Conected FANN with Condensed Information and Modulation
$SAFC$	Shuffle Architecture of FC FANN
$SAMFC$	Shuffle Architecture of FC FANN with Modulation
$SACIFC$	Shuffle Architecture of FC FANN with Condensed Information
$SACIMFC$	Shuffle Architecture of FC FANN with Condensed Information and Modulation

4.6 Tabular Methods

Some special considerations are required when using tabular methods. In such techniques, there are no input neurons to receive the chain state directly. Instead, all possible state combinations must be represented in a table (the Q-values table), which stores the expected accumulated reward for each state-action pair. This table is then used to guide the action selection policy.
Consequently, we need to store a value for each pair (s, a), where s is the environment state and a is the action to be executed. The set of actions is known to be finite; specifically, we consider n actions, where n is the length of the chain.

As mentioned earlier, we adopt the **CI** approach for tabular methods. In this representation, the state is defined as a pair (p, a), where $p \in \mathbb{I}$ is an integer denoting the position of the wave peak (with $0 \leq p \leq n-1$), and $a \in \mathbb{R}$ is a real number representing the wave peak amplitude.

Since a is a continuous value, it cannot be used directly in a tabular representation. Therefore, we discretize the amplitude to define the environment states as elements of the Cartesian product $\mathbb{P} \times \mathbb{A}$, where $\mathbb{P}$ is the set of possible positions and $\mathbb{A}$ is the set of discretized amplitude levels. We chose to define 15 discrete amplitude levels and map the actual amplitude from the environment to the nearest among these 15 levels.

To clarify, if we are solving the problem for a chain of length n, the resulting Q-value table will have a size of $n \times 15 \times n$.

Table 5. FANN's dynamical configuration. For each architecture number of neurons and connection are shown grouped by type. *np* and *cp* refer to total neuron and connection parameters, respectively.

No length dependant											
	neurons				connections				dyn param		
	in	out	intern.	total	inhib.	excit.	elctr.	total	np	cp	total
TWC,	4	2	5	11	15	9	2	26	22	26	48
MTWC	4	2	5	11	15	9	2	26	22	52	74
CIFC	2	1	7	10	12	7	2	21	20	21	41
MCIFC	2	1	7	10	12	7	2	21	20	42	62
Length dependant											
Qbit7											
FC	7	1	9	17	40	25	7	72	34	72	106
MFC	7	1	9	17	40	25	7	72	34	144	178
Qbit10											
FC	10	1	15	26	91	58	16	165	52	165	217
MFC	10	1	15	26	91	58	16	165	52	330	382
Qbit13											
FC	13	1	21	35	162	103	29	294	70	294	364
MFC	13	1	21	35	162	103	29	294	70	588	658
Qbit16											
FC	16	1	27	44	252	161	46	459	88	459	547
MFC	16	1	27	44	252	161	46	459	88	918	1006

4.7 Experimental Setup

In this section, we present the experiments conducted to evaluate the performance of various combinations applied to solve the four problem instances defined as **Qbit7**, **Qbit10**, **Qbit13**, and **Qbit16**. A total of 3,600 models were trained, corresponding to 900 models per problem instance.

For each problem, we explored the four reward versions introduced earlier. Thus, for each combination of problem *Qbitn* and reward function r, we trained 225 models.

This number results from the following configuration space: we trained 12 neural architectures (listed in Table 4), each with one of three optimization algorithms—**RS** (Random Search), **GA** (Genetic Algorithm), and **BO** (Bayesian Optimization)—resulting in 36 different **FANN** configurations. In addition, we trained three tabular methods, each with three different step sizes, adding 9 tabular configurations. This gives a total of 45 configurations per $(Qbitn, r)$ pair.

Each configuration was trained and evaluated five times to account for variability, resulting in $45 \times 5 = 225$ models per pair. Therefore, across the 4 problem

instances and 4 reward functions, a total of $4 \times 4 \times 225 = 3,600$ models were trained and analyzed.

FANN models incorporate an early detection mechanism (Early Stopping), so in all cases, a large number of optimization steps was set. However, the specific number of steps varies according to the optimization algorithm used: for **RS**, 500,000 steps were used; for **GA**, 5,000; and for **BO**, 10,000. All **FANN** models were trained using the number of steps corresponding to the optimizer used.

Instead, tabular methods do not have an early detection mechanism. For this reason, three scenarios per configuration were evaluated, using 100,000, 200,000, and 400,000 episodes, respectively. In this case, we talk about episodes and not optimization steps.

Once a model was trained, it underwent an evaluation phase. This consisted of attempting to solve the problem again using the trained model, starting from the initial state and measuring the final transmission fidelity. The resulting fidelity is reported as the performance metric in this work. Since the problem is deterministic, a single evaluation run is sufficient. By last for each chain length, we took three non-tabular models appearing at a distance less than 5% of the best one and trained them 10 times. Also trained 10 times tabular models to compare them in a more populated statistical scenario.

5 Results

In this section, we report the results obtained and compare them with related works. The results, presented in Tables 6 to 8, show that our study not only stands out in terms of absolute fidelity and stability, but also exceeds the scope of previous work in the literature, particularly that of Zhang et al. [17].

Table 6. Experimental results disaggregated by reward type, some models have in their names the string **RS**, **GA** or **BO** only to remark the optimization method used to reach that solution.

Reward	Fid	Model	Reward	Fid	Model
Qbit7			Qbit10		
LastAuto.	0.993	MC	*LastAuto.*	0.987	MC
RandomAuto.	0.975	$CIFC_{GA}$	*RandomAuto.*	0.951	$MSATWC_{BO}$
SiteEvol.	0.994	MFC_{RS}	*SiteEvol.*	0.913	$SAMFC_{GA}$
Transm.Evol.	0.993	*DoubleQN*	*Transm.Evol.*	0.876	$SAMFC_{BO}$
Qbit13			Qbit16		
LastAuto.	0.967	MC	*LastAuto.*	0.937	MC
RandomAuto.	0.928	$CIMFC_{GA}$	*RandomAuto.*	0.892	MFC_{BO}
SiteEvol.	0.846	$SACIMFC_{BO}$	*SiteEvol.*	0.854	$SAFC_{GA}$
Transm.Evol.	0.828	TWC_{GA}	*Transm.Evol.*	0.925	FC_{RS}

Table 7. Experimental results disaggregated by model type, some models have in their names the string **RS**, **GA**, or **BO** only to remark the optimization method used to reach that solution. Every problem is highlighted with dark gray the best obtained value($\mathcal{B}$), with gray a value whose distance $\mathcal{D}$ to $\mathcal{B}$ is less than $\mathcal{B} * 0.01$, and with light gray the values with $\mathcal{B} * 0.01 \leq \mathcal{D} \leq \mathcal{B} * 0.05$. To allow some additional analysis, we also show in the *ip* column the number of internal parameters in the case of FANNs and results obtained running the code coming from Zhang et al. [17].

Archit	Fid	Model	ip	Archit	Fid	Model	ip
Qbit7				Qbit10			
TWC	0.978	TWC_{GA}	48	TWC	0.936	TWC_{BO}	48
$SATWC$	0.963	$SATWC_{RS}$	48	$SATWC$	0.925	$SATWC_{GA}$	48
$MTWC$	0.964	$MTWC_{BO}$	74	$MTWC$	0.931	$MTWC_{RS}$	74
$MSATWC$	0.973	$MSATWC_{RS}$	74	$MSATWC$	0.951	$MSATWC_{BO}$	74
FC	0.98	FC_{GA}	106	FC	0.949	FC_{GA}	106
MFC	0.994	MFC_{RS}	178	MFC	0.96	MFC_{GA}	178
$CIFC$	0.975	$CIFC_{GA}$	41	$CIFC$	0.946	$CIFC_{GA}$	41
$CIMFC$	0.973	$CIMFC_{BO}$	62	$CIMFC$	0.913	$CIMFC_{BO}$	62
$SAFC$	0.981	$SAFC_{BO}$	106	$SAFC$	0.96	$SAFC_{GA}$	106
$SAMFC$	0.965	$SAMFC_{GA}$	178	$SAMFC$	0.953	$SAMFC_{GA}$	178
$SACIFC$	0.954	$SACIFC_{BO}$	41	$SACIFC$	0.897	$SACIFC_{GA}$	41
$SACIMFC$	0.919	$SACIMF_{BO}$	62	$SACIMFC$	0.89	$SACIMFC_{BO}$	62
QN	0.992	QN	–	QN	0.829	QN	–
$DoubleQN$	0.993	$DoubleQN$	–	$DoubleQN$	0.946	$DoubleQN$	–
MC	0.993	MC	–	MC	0.987	MC	–
$Zhang$	0.999	–	7296	$Zhang$	0.977	–	8016
Qbit13				Qbit16			
TWC	0.909	TWC_{BO}	48	TWC	0.907	TWC_{BO}	48
$SATWC$	0.895	$SATWC_{GA}$	48	$SATWC$	0.883	$SATWC_{BO}$	48
$MTWC$	0.885	$MTWC_{BO}$	74	$MTWC$	0.901	$MTWC_{RS}$	74
$MSATWC$	0.91	$MSATWC_{GA}$	74	$MSATWC$	0.872	$MSATWC_{BO}$	74
FC	0.939	FC_{GA}	106	FC	0.93	FC_{RS}	106
MFC	0.917	MFC_{GA}	178	MFC	0.892	MFC_{BO}	178
$CIFC$	0.919	$CIFC_{GA}$	41	$CIFC$	0.921	$CIFC_{BO}$	41
$CIMFC$	0.928	$CIMFC_{GA}$	62	$CIMFC$	0.849	$CIMFC_{BO}$	62
$SAFC$	0.945	$SAFC_{GA}$	106	$SAFC$	0.909	$SAFC_{GA}$	106
$SAMFC$	0.916	$SAMFC_{GA}$	178	$SAMFC$	0.883	$SAMFC_{BO}$	178
$SACIFC$	0.908	$SACIFC_{BO}$	41	$SACIFC$	0.85	$SACIFC_{RS}$	41
$SACIMFC$	0.884	$SACIMFC_{BO}$	62	$SACIMFC$	0.814	$SACIMFC_{BO}$	62
QN	0.739	QN	–	QN	0.889	QN	–
$DoubleQN$	0.922	$DoubleQN$	–	$DoubleQN$	0.918	$DoubleQN$	–
MC	0.967	MC	–	MC	0.937	MC	–
$Zhang$	0.966	–	8736	$Zhang$	0.943	–	9456

Experimental results show that multiple control strategies achieve high fidelities for the quantum transmission task. In particular, the tabular method **Monte Carlo (MC)**, in most cases, was the most efficient in terms of average fidelity across different chain lengths (see Table 7). These approaches achieve results above 0.93 even on chains of up to 16 qubits, and they do so without requiring complex neural structures, as they store estimates per state or action. Even

Table 8. Experimental results disaggregated by model type, here we took the three best non tabular models and also the tabular ones. All that models were trained ten times and gathered the maximum fidelity, the mean and the standar variation with respect to that ten trainings. In that table also are the reference to Zhang et al. [17] trained ten times. Every problem is highlighted with dark gray the best obtained value($\mathcal{B}$), with gray a value whose distance $\mathcal{D}$ to $\mathcal{B}$ is less than $\mathcal{B} * 0.01$, and with light gray the values with $\mathcal{B} * 0.01 \leq \mathcal{D} \leq \mathcal{B} * 0.05$.

Model	Max	Mean	Std	Model	Max	Mean	Std
Qbit7				Qbit10			
MFC	0.994	0.834	0.156	*MFC*	0.960	0.928	0.013
SAFC	0.971	0.954	0.016	*SAFC*	0.967	0.939	0.016
FC	0.980	0.971	0.007	*SAMFC*	0.953	0.927	0.015
QN	0.992	0.783	0.245	*QN*	0.837	0.661	0.171
DoubleQN	0.993	0.954	0.049	*DoubleQN*	0.946	0.826	0.178
MC	0.994	0.979	0.025	*MC*	0.987	0.901	0.038
Zhang	0.999	0.998	0.001	*Zhang*	0.984	0.975	0.005
Qbit13				Qbit16			
SAFC	0.945	0.914	0.015	*FC*	0.930	0.901	0.021
FC	0.946	0.917	0.018	*SAFC*	0.942	0.890	0.021
CIMFC	0.928	0.870	0.034	*TWC*	0.907	0.872	0.025
QN	0.799	0.436	0.187	*QN*	0.912	0.746	0.025
DoubleQN	0.952	0.915	0.034	*DoubleQN*	0.918	0.894	0.018
MC	0.967	0.946	0.029	*MC*	0.937	0.930	0.004
Zhang	0.967	0.960	0.006	*Zhang*	0.946	0.937	0.006

more, taking into account the analysis based on a few models with 10 trainings (Table 8), we can see that bio-inspired models are very close (in one case even better than) reference results of Zhang et al. [17].

However, the advantage of tabular methods is not overwhelming: several neural network-based models achieved similar fidelities. For example, the fully connected architecture (**FC**) and its modular version (**MFC**) or even **SA**'s also achieved high reliability. However, the complexity of these networks scales significantly with the number of qubits. According to Table 5, for 16 qubits, an **MFC** network exceeds 900 trainable parameters, which implies a substantially higher computational cost. In this context, pure bio-inspired architectures stand out, especially the **TWC**, which achieves very competitive fidelities (surpassing 0.90) with just 11 neurons and about 40 estimable parameters.

These architectures do not depend on chain length and maintain their size constant even when scaling the system, as seen in the upper part of Table 5. This makes them ideal candidates for scenarios where structural robustness and implementation efficiency are required. All these conclusions are reinforced by the

results obtained on the 10 training models. Those results are shown in Table 8. Here, we observe that the performance gap between bio-inspired models and tabular ones narrows for two problems (Qbit10 and Qbit13), while for Qbit7 the results are identical. On the other hand, the best outcome for Qbit16 is achieved by a bio-inspired model (**SAFC**). Finally, when comparing against our reference baseline, we find that the 10 trained models also reduce the gap with respect to Zhang et al. [17] Compared with previous work, our approach has clear advantages. We successfully reproduced the results of Zhang et al.'s work in the evaluation phase, achieving similar fidelities on chains of up to 11 qubits. However, there are fundamental differences in the methodological approach. Zhang et al.'s work does not seek to train a generalizing **RL** agent, but rather to optimize a policy on a validation set under ideal conditions. Their implementation used fully connected networks of more than 200 neurons, exceeding 7,000 trainable parameters. In contrast, our models achieve similar or higher fidelities using two to three orders of magnitude fewer parameters.

6 Conclusion and Future Work

This study demonstrates that it is possible to design quantum control agents that are efficient in both fidelity and structural complexity. Tabular methods such as **MC** and **Double Q-learning** emerged as the most effective, although architectures such as **TWC** and **FC** achieved comparable fidelities, with varying degrees of computational cost. In particular, **TWC** achieved fidelities above 0.9 using only a fraction of the parameters required by traditional fully connected networks. This result validates the hypothesis that performance does not depend exclusively on the number of parameters, but on the agentn's structural design.

Bio-inspired architectures, by incorporating functional motifs such as synaptic modulation and distinct neuronal types (inhibitory, excitatory, electrical), were able to maintain high performance with fixed, compact, and scalable topologies. This represents a promising alternative to dense neural networks, especially for applications on limited hardware or real quantum systems.

Unlike the work of Zhang et al. [17], whose main objective was to validate effective protocols without training generalizable agents, our approach aims to develop truly autonomous agents. Although in this work the agent is deterministic (due to the characteristics of the environment), the future objective is to extend training to non-deterministic environments. To this end, we propose:

- **Incorporate structured and random noise during training** to robustify the learned policy against realistic perturbations and errors.
- **Develop adaptive agents** that do not require ex post selection of the best policy, but rather converge through experience and reinforcement.
- **Explore novel bio-inspired dynamics**, such as rhythmic oscillations or recurrent couplings, which could improve coordination in information transfer along chains of qubits.
- **Evaluate the combination of strategies**, integrating tabular methods with compact neural modules, to take advantage of the best of both worlds.

– **Scalability study**, developing methods that generalize the chain length or increase it with a manageable cost.

Taken together, these results suggest that the use of neurobiologically inspired functional structures, combined with robust reinforcement learning techniques, can lead to the development of scalable, efficient, and adaptable quantum controllers suitable for applications in experimental physics and emerging technologies. Preliminary results show that deep learning strategies require networks with a very large number of neurons which are computationally demanding to train [11]. Moreover, there is no indication that changing the reward or the ϵ-greedy policy for other ones results in better training or more robust models. The higher values of the transmission probabilities obtained with genetic algorithms [11] inform us that both, deep learning strategies could lack the exploring abilities needed to find better policies on the parameter space and that robust control strategies do exists when the control pulses show low-enough fluctuations. Although there is no a clear-cut relationship between control of quantum mechanical problems and biologically motivated networks, using the latter ones open a different path, employing networks with much less parameters with better scaling properties and more efficient learning strategies.

References

1. Abbott, L.F., Dayan, P.: Theoretical neuroscience (2021). Accessed 30 July 2025
2. Acosta Coden, D.S., Gómez, S.S., Ferrón, A., Osenda, O.: Controlled quantum state transfer in xx spin chains at the quantum speed limit. Phys. Lett. A **387**, 127009 (2021)
3. Fiuri, A.M., Dominguez, M.A., Tamarit, F.: Reinforcement learning and biologically inspired artificial neural networks. In: Information Management and Big Data, pp. 62–79. Springer, Cham (2024)
4. Hall, D., Chen, B., Chklovskii, D.: Wiring optimization can relate neuronal structure and function. Proc. Natl. Acad. Sci. **103**(12), 4723–4728 (2006)
5. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 17–19 June 2013, vol. 28, pp. 115–123. PMLR (2013)
6. Brochu, E., Cora, V.M., de Freitas, N.: A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning (2010)
7. de Keijzer, R.J.P.T., Visser, L.Y., Tse, O., Kokkelmans, S.J.J.M.F.: Fidelity-enhanced variational quantum optimal control. Phys. Rev. A **111**, 052625 (2025)
8. Gad, A.F.: Pygad: an intuitive genetic algorithm python library (2023)
9. Gong, J., Brumer, P.: Controlled quantum-state transfer in a spin chain. Phys. Rev. A **75**, 032331 (2007)
10. Murphy, M., Montangero, S., Giovannetti, V., Calarco, T.: Communication at the quantum speed limit along a spin chain. Phys. Rev. A **82**, 022318 (2010)
11. Santana, S.P., Fiuri, A., Domínguez, M., Osenda, O.: The pros and cons of using deep reinforced learning or genetic algorithms to design control schemes for quantum state transfer on qubit chains. in preparation (2025)

12. Ghashghaei, F.R., Elmrabit, N., Qureshi, A.U.H., Akhunzada, A., Yousefi, M.: Advanced quantum control with ensemble reinforcement learning: a case study on the xy spin chain. IEEE Access **13**, 49514–49526 (2025)
13. Santana, S.P., Fiuri, A., Osenda, O., Domínguez, M.: Optimización de la transmisión de estados cuánticos en cadenas de qubits usando deep reinforcement learning y algoritmos genéticos, arxiv: 2510.05010 (2025)
14. Sgroi, S., Zicari, G., Imparato, A., Paternostro, M.: A reinforcement learning approach to the design of quantum chains for optimal energy and state transfer. Mach. Learn. Sci. Technol. **6**(1), 015012 (2025)
15. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. IEEE Trans. Neural Netw. **9**(5), 1054–1054 (1998)
16. White, J.G., Southgate, E., Thomson, J.N., Brenner, S.: The structure of the nervous system of the nematodecaenorhabditis elegans. Philos. Trans. R. Soc. Lond. **314**(1165), 1–340 (1986)
17. Zhang, X.-M., Cui, Z.-W., Wang, X., Yung, M.-H.: Automatic spin-chain learning to explore the quantum speed limit. Phys. Rev. A **97**, 052333 (2018)

Horoyah: A Knowledge Graph-Based Framework for Election Monitoring in Low and Middle-Income Countries

Gayo Diallo(✉)

BPH Inserm U1219, Univ. Bordeaux, 33000 Bordeaux, France
gayo.diallo@u-bordeaux.fr

Abstract. In many low and middle-income countries (LMICs), electoral processes are plagued by dysfunctions that undermine democratic principles and public trust. These issues, ranging from political corruption to a lack of transparency, often lead to post-election violence and contested results. Civil society organizations (CSOs) play a crucial role in mitigating these problems through election monitoring, but often lack the tools for effective, large-scale data collection and analysis. This paper introduces Horoyah, an innovative platform designed to empower civil society and the general public by leveraging knowledge graphs to promote transparency and accountability in election monitoring. Horoyah provides a structured framework for collecting, aggregating, and analysing citizen-reported election results from polling stations. Structuring voting data within a knowledge graph enables the platform to facilitate the identification of trends, inconsistencies, and potential irregularities, thereby enhancing data accuracy and providing an auditable record of the electoral process. We outline its design and demonstrate its application through a use case based on the 2020 presidential elections in Guinea, West Africa. We conclude by discussing the potential of such tools to fortify democratic practices while also considering the significant challenges of disinformation, cybersecurity, and the digital divide that they raise.

Keywords: Election Monitoring · Knowledge Graph · Civil Society · Transparency · LMIC · e-Democracy

1 Introduction

The wave of democratization that swept across Africa in the early 1990s transitioned many nations from single-party rule to multi-party systems. Despite significant progress in the practice of holding elections, many have devolved into mere formalities that serve to approve political elites, undermining the credibility of the entire process. In many African Low and Middle Income Countries (LMICs) like Guinea, Kenya, and Zimbabwe, electoral processes are frequently marred by claims of vote-rigging and subsequent violence. These dysfunctions

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 79–92, 2026.
https://doi.org/10.1007/978-3-032-20322-9_6

permeate nearly every stage of the process, from the formation of politically compromised electoral bodies to the announcement of results, creating an environment ripe for irregularities. This systemic lack of transparency erodes voters' trust and makes the ideals of democracy unattainable.

Information and Communication Technology (ICT) has been investigated as a means to implement e-democracy and combat corruption, with evidence suggesting that e-government systems can increase transparency [2]. Empirical studies have shown an inverse relationship between ICT usage and perceived corruption, highlighting technology's potential as a valuable tool in anti-corruption efforts [5].

In this context, Civil Society Organizations (CSOs) are critical actors in promoting electoral transparency [9]. Through independent observation and public awareness campaigns, CSOs contribute significantly to the credibility and integrity of elections [12]. The active engagement of citizens in monitoring elections is a cornerstone of a robust democracy, and technology offers a transformative potential to bolster this civic role [13].

This paper introduces Horoyah, a platform engineered to harness the power of knowledge graphs to advance election monitoring for civil society in LMICs. A knowledge graph is a structured data representation that organizes information and establishes connections between different data points [6]. Applying this to voting data allows for a more comprehensive analysis of the relationships between voter demographics, voting patterns, and results. This method enhances data accuracy, enables more efficient analysis, and promotes transparency by providing a clear, auditable trail. We present its overall design and implementation and a use case based on the 2020 Guinean presidential election. We argue that by providing tools for citizen-led data collection and analysis, Horoyah can empower CSOs and the public, fortifying democratic electoral practices in resource-constrained environments.

2 Background

The field of e-democracy has explored various technological interventions to improve electoral processes. Early studies focused on electronic polling services for creating polls and viewing results [2], while others advocated for e-democracy tools like "smart vote" to enhance democratic principles [8]. A significant body of research has been dedicated to internet voting, examining its advantages and challenges [12], and proposing secure, end-to-end verifiable systems like D-DEMOS [3]. In Estonia, a nationwide e-Election System has been implemented, allowing for online voting[1]. These efforts highlight the long-standing interest in using ICT to improve the efficiency and transparency of elections.

Another stream of work focuses on making election data more accessible and usable. In the United States, for instance, the OpenElections project, launched in 2012, addresses the challenge of obtaining comprehensive and standardized

[1] https://digiexpo.e-estonia.com/e-governance/e-democracy-and-open-data/e-election-system/.

election data[2]. It utilizes a network of volunteers to collect and convert official results into standardized CSV formats, which are then shared on GitHub. This model has proven valuable for journalists and researchers. Similarly, the electiondata Python package was developed to help users consolidate, analyze, and visualize election results from various file formats [13].

In the context of LMICs, specific technological solutions have been proposed to address local challenges. In India for instance, an "Integrated Election Voting System" model was proposed to supplement existing Electronic Voting Machines (EVMs) with an e-voting channel to increase participation [9]. The study emphasized the potential for ICT to enhance India's electoral process and increase voter turnout. In Senegal, civil society initiatives like "Sénégal Ouvert" and open data repositories on GitHub demonstrate a growing movement towards election transparency through ICT technology[3].

One innovative technological approach is the use of knowledge graphs for the collection and aggregation of voting data. The novelty of Horoyah includes the application of knowledge graphs to the voting domain. A knowledge graph offers a structured way to link diverse data points like voter demographics and voting patterns, enabling deeper analysis. This approach enhances data accuracy, facilitates efficient processing, and provides an auditable trail of data, which is crucial for transparency. Research has explored using networks-based approach to improve problems of poor service delivery and quality of care in low-income regions, aiming to foster more inclusive approaches [7]. Horoyah builds on these ideas to create a concrete, deployable platform specifically for civil society-led election monitoring in LMICs.

3 Overall Description of the Platform

3.1 Functional Components

The Horoyah platform is comprised of several key components that work together to create a seamless workflow for election monitoring.

Data Collection. Horoyah is designed for ease of use by citizen reporters at polling stations. A user-friendly, location-based smart form, similar to applications like Survey, is used for data entry. This allows observers to efficiently submit their observations, including vote counts and reports of any incidents.

Data Aggregation and Verification. Once submitted, the data is integrated into the Horoyah knowledge graph (please see Fig. 1). A critical step in this process is verification. To maintain the integrity of the information, the W3C Shapes Constraint Language (SHACL)[4] is used for validation, complemented by a team of volunteers screening each submission before it is taken into account within the knowledge graph and included on a public map or dashboard.

[2] https://github.com/openelections.
[3] https://github.com/senegalouvert.
[4] https://www.w3.org/TR/shacl/.

Analysis and Visualization. The true power of the knowledge graph lies in the analytical capabilities that it can enable. Linking voting data with demographic information and geographic locations enables the possibility to identify correlations and trends within the electorate, such as patterns based on age, gender, or socioeconomic status. Further, a crosscheck with the available national census and other demographic characteristics openly available could be possible. The platform can then visualize this enriched data through dashboards and interactive maps as shown in Fig. 3, allowing the public and media to easily explore the results and focus on specific issues or areas.

Knowledge Graph Population. The verified data populates the ontology schema shown in Fig. 1, creating a rich, interconnected dataset. This process automates and streamlines what would otherwise be a manual effort prone to errors and delays. The structured nature of the knowledge graph enhances data accuracy by minimizing inconsistencies during collection and aggregation.

Cartographic Data. OpenStreetMap is used for the maps and geographical structures description[5] [4] (but could also be LinkedGeoData[6] [1]). The OpenStreetMap platform is built by a community of volunteer cartographers who

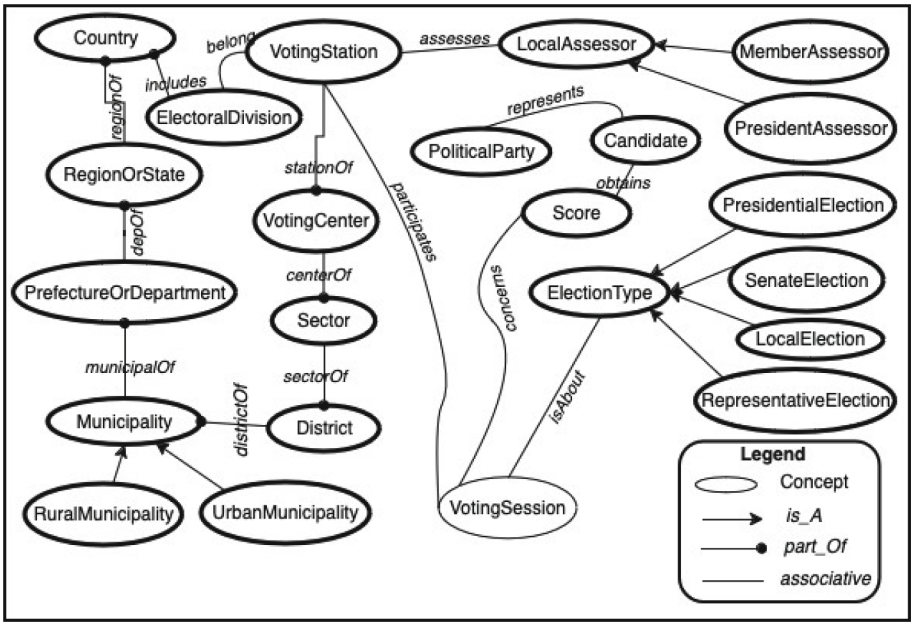


Fig. 1. The Horoyah Knowledge Graph model used in the platform.

[5] https://www.openstreetmap.org.
[6] https://linkedgeodata.org/.

contribute and maintain data on neighbourhoods, roads, trails, cafés, train stations and much more, all over the world. It is therefore an open, collaborative map of the world. It is improved every day by over a million contributors.

3.2 The Horoyah Knowledge Graph

The core component of Horoyah data and knowledge backbone is modelled as a knowledge graph, representing the various entities and their complex relationships within an electoral system. Figure 1 illustrates the ontology model of the knowledge graph. The core entities include the following notions.

- **Administrative and Electoral Divisions**: The model captures the hierarchical structure of a country, from the national level down to RegionOrState, PrefectureOrDepartment, Municipality, and District.
- **Voting Infrastructure**: This includes VotingCenter and VotingStation, which are the primary locations where voting occurs and results are reported from. The model can be enhanced with geographical data from open sources like OpenStreetMap [4], which provides a collaborative map of the world built by volunteers. It could also be any other geographical data framework such as LinkedGeoData [1].
- **Electoral Process**: Key entities like ElectionType (e.g., PresidentialElection), PoliticalParty, and Candidate are central to the model. The VotingSession concept links these entities for a specific election event.
- **Actors and Results**: LocalAssessor represents the civil society observer or citizen who assesses the process and reports the Score (results) obtained by a Candidate at a VotingStation.

This interconnected structure allows the platform to link disparate pieces of information, such as results from a specific polling station, the political party of the candidates, and the demographic data of the electoral subdivisions (Fig. 2).

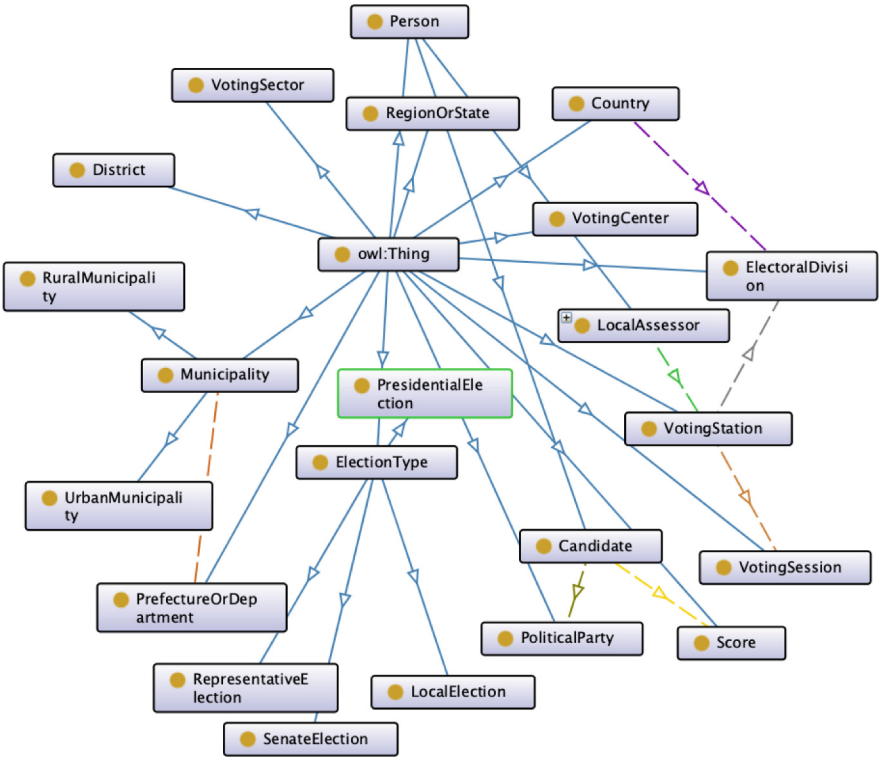


Fig. 2. Screenshot of the graph rendered with the OntoGraph panel of the Protégé editor tool.

3.3 Validating the Knowledge Graph

SHACL is a language for validating Resource Description Framework (RDF) based graphs against a set of conditions [11]. These conditions are provided as "shapes" and other constructs expressed in the form of an RDF graph. SHACL is a W3C Consortium recommendation that helps ensure data quality and consistency within a knowledge graph. Three examples are provided in the following listings, respectively for the CandidateShape, Subclass Shapes (ElectionType-SubClassShape) and *participatesIn* property.

CandidateShape

```
@prefix oekg: <http://www.horoya.org/OpeneElectionKG#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .

oekg:CandidateShape
    a sh:NodeShape ;
    sh:targetClass oekg:Candidate ;
    sh:property [
        sh:path oekg:obtains ;
```

```
        sh:class oekg:Score ;
        sh:name "obtains" ;
        sh:description "A candidate must obtain a score." ;
        sh:minCount 1 ;
    ] ;
    sh:property [
        sh:path oekg:reprsents ;
        sh:class oekg:PoliticalParty ;
        sh:name "represents" ;
        sh:description "A candidate must represent a
    political party." ;
        sh:minCount 1 ;
    ] .
```

Subclass Shapes. These shapes ensure that the ElectionType sub-classes axioms from the ontology are respected.

```
@prefix oekg: <http://www.horoya.org/OpeneElectionKG#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

oekg:ElectionTypeSubClassShape
    a sh:NodeShape ;
    sh:targetClass oekg:LocalElection, oekg:
    PresidentialElection, oekg:RepresentativeElection, oekg:
    SenateElection ;
    sh:property [
        sh:path rdf:type ;
        sh:hasValue oekg:ElectionType .
    ] .
```

Rules for *participatesIn* Property. The *participatesIn* property links a *VotingStation* to a *VotingSession*. Here are some rules to validate this relationship: *sh* : *pathopenelectionKG* : *participatesIn*. This rule specifies the *participatesIn* property as the one to be validated. Here are some rules to validate this relationship.

- sh:path openelectionKG:participatesIn: This rule specifies the participatesIn property as the one to be validated.
- sh:class openelectionKG:VotingStation: This rule specifies that the subject of the participatesIn property must be an instance of the VotingStation class.
- sh:targetClass openelectionKG:VotingSession: This rule asserts that the object of the participatesIn property must be an instance of the VotingSession class.

```
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix oekg: <http://www.horoya.org/OpeneElectionKG#> .

# Shape to validate the 'participatesIn' property of a VotingStation

oekg:VotingStationShape
    a sh:NodeShape ;
    sh:targetClass oekg:VotingStation ;

    # Defines the property constraint
    sh:property [
        # This is the core of the rule, specifying the property to be validated.
        sh:path oekg:participatesIn ;

        # Human-readable name for the rule
        sh:name "Voting Station Participation" ;

        # Description of what the rule checks
        sh:description "Validates the participatesIn relationship for a VotingStation." ;
    ] .
```

3.4 Querying the Knowledge Graph

The Horoyah Knowledge graph is expressed in RDF using the triplet based representation. It can be querying therefore using the SPARQL query language. We provide in the following an example of a SPARQL query that calculates the total vote score for a given candidate from a specific political party in a given region of the country during a presidential election. The SELECT clause [... (SUM(?scoreValue) AS ?totalScoreInRegion)] selects the names of the candidate, party, and region, and most importantly, it calculates the sum (SUM) of all found score values (?scoreValue). It renames this total as ?totalScoreInRegion for clarity.

```
    # Prefixes for readability
PREFIX oekg: <http://www.horoya.org/OpeneElectionKG#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

# Query to aggregate scores
SELECT ?candidateName ?partyName ?regionName (SUM(?scoreValue) AS ?totalScoreInRegion)
```

```
WHERE {
    # 1. Identify the election type and the associated voting session
    ?session rdf:type oekg:VotingSession .
    # Assumption 1: The session is linked to an election type
    ?session oekg:isForElectionType oekg:PresidentialElection .

    # 2. Identify the candidate and the political party
    ?candidate rdf:type oekg:Candidate .
    ?candidate oekg:repersents ?party .

    # 3. Retrieve the candidate's score for this session
    ?candidate oekg:obtains ?score .
    # Assumption 2: The score is linked to a specific voting session
    ?score oekg:inSession ?session .
    # Assumption 3: The score has a numeric value
    ?score oekg:hasValue ?scoreValue .

    # 4. Link the vote to its geographical location (the Region)
    ?station oekg:participatesIn ?session .
    ?station oekg:belongTo ?electoralDivision .
    # Assumption 4: The electoral division is located in a region
    ?electoralDivision oekg:locatedIn ?region .
    ?region rdf:type oekg:RegionOrState .

    # 5. Retrieve names for display (Assumption 5)
    ?candidate rdfs:label ?candidateName .
    ?party rdfs:label ?partyName .
    ?region rdfs:label ?regionName .

    # 6. Filter for the specified party and region
    FILTER(?partyName = "GivenPartyName"@en)
    FILTER(?regionName = "GivenRegionName"@en)
}
# 7. Group the results to sum by candidate/party/region
GROUP BY ?candidateName ?partyName ?regionName
# 8. Order the results
ORDER BY DESC(?totalScoreInRegion)
```

4 Illustration: the Guinea's 2020 Presidential Elections

4.1 Context of the Presidential Elections

Guinea is a republic located in the West African region and a member of the ECOWAS regional organization. A Presidential elections were held on 18 October 2020, with incumbent president Alpha Condé (RPG party) running for a third term against former prime minister Cellou Dalein Diallo (UFDG party) and various other candidates. Earlier in that year, a controversial constitutional referendum reset presidential term limits, allowing Condé to run again. Despite opposition boycotts and protests against Condé's government, the elections went ahead alongside a legislative vote. The electoral process, overseen by the independent electoral commission (CENI), used a two-round system where the President is elected by an absolute majority vote for a 5-year term. The country's population during the 2020 election was reported as 12,527,440, with an estimated 5,179,600 registered voters in 2020[7]. Initial reports indicated that Diallo was in the lead, and on 19 October, he declared victory prematurely based on his party's polling, which was criticized by Condé. However, with 96.14% of votes counted, the national election committee (CENI) officially announced Condé as the winner in the first round with 59.49% of the vote on 24 October. During the voting date and the following days, the Union of Free Radios and Televisions of Guinea (URTELGUI), an independent SCO monitored around 11% of the voting stations nationally, which involved 250 journalists. A dedicated radio broadcast was used to report the results. We used the available dataset to instantiate Horoyah. Table 1 summarizes the official results as provided by the CENI. These results were strongly contested by the 2nd place candidate and his party. As could be seen, SCO based results, as depicted in Fig. 3 suggest a noticeable difference with those provided by the official CENI committee. This raises the issue of the transparency during the official aggregation and counting process of the election results.

Table 1. Results of the 2020 presidential elections as proclaimed by the CENI

Candidate	Political Part	Votes obtained	Percentage
Alpha Condé	RPG	2,438,815	59.49
Cellou Dalein Diallo	UFDG	1,373,320	33.50
Ibrahima Abé Sylla	NGR	63,676	1.55
Ousmane Kaba	PADES	48,623	1.19

5 Discussions

The integration of digital tools like Horoyah into election monitoring presents both profound opportunities and significant challenges. While they can enhance

[7] https://www.idea.int/data-tools/country-view/115/40.

transparency and empower citizens and CSO, they also introduce vulnerabilities that must be carefully managed to safeguard electoral integrity.

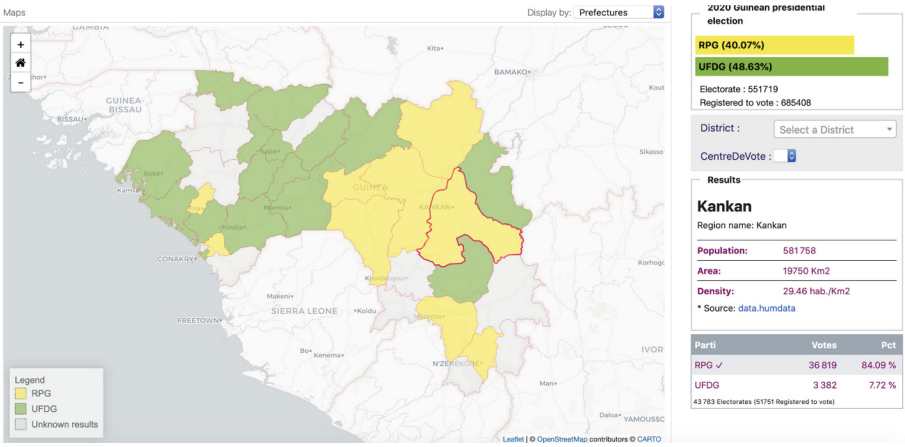


Fig. 3. Observed results nationwide by the URTELGUI SCO, with an highlight for Kankan, a north-east region in Guinea. The yellow colour if for the candidate of the RPG party while the green colour for the UFDG. (Color figure online)

5.1 Benefits and Opportunities

The primary benefit of a framework like Horoyah is the enhancement of transparency and public trust. By providing a transparent avenue for reporting and addressing issues, the system promotes accountability. Online platforms also improve accessibility, allowing voters to participate in the democratic process from various locations, which can lead to increased voter participation, especially among younger, tech-savvy demographics. Furthermore, transitioning from paper-based reporting to digital systems contributes to environmental sustainability by reducing paper waste and the carbon footprint associated with elections.

5.2 Challenges and Limitations

Despite the benefits, the deployment of such technologies is fraught with challenges. We highlight in the following sections some of them.

Disinformation and Cybersecurity. A major concern is the risk of disinformation. Malicious actors could attempt to undermine the system by flooding it with false reports. This necessitates robust verification mechanisms and collaboration between civil society and media to counteract misleading information.

Cybersecurity is another pressing issue. Online systems are vulnerable to cyber threats, including denial-of-service (DoS) attacks that could disrupt the reporting process, and hacking that could compromise the privacy and anonymity of reporters.

The Digital Divide. In many LMICs, limited internet access and varying levels of digital literacy pose significant barriers to participation. A digital-only platform risks excluding large segments of the population, particularly in rural areas. Any successful implementation must consider context-sensitive approaches, potentially integrating with lower-tech solutions like SMS or providing offline capabilities.

Platform Integrity. Digital platforms like WhatsApp and Telegram have been used in countries like Brazil and India to disseminate misinformation and coordinate attacks on the electoral process. This highlights that the platforms themselves can become vectors for undermining electoral integrity, often lacking robust policies to operate safely in diverse linguistic and cultural contexts[8].

5.3 The Role of AI and Future Trends

The rise of Artificial Intelligence (AI) presents a new frontier of challenges and opportunities for elections. Generative AI can be used to create highly realistic "deepfakes" and spread disinformation on an unprecedented scale, posing a critical risk to election integrity[9]. Lawmakers and civil society must prepare to address these evolving threats.

However, AI also offers powerful tools that can be harnessed for good. AI algorithms could be integrated into Horoyah to perform advanced analysis on incoming data, automatically detecting anomalies and flagging potential irregularities for review. As envisioned in the project's future work, AI could also power voice-based applications for low-literacy users [10], making the reporting process more inclusive. The key will be to develop these technologies responsibly, with a focus on augmenting human oversight rather than replacing it, to ensure they strengthen rather than subvert democratic processes.

6 Conclusion and Perspectives

This paper has introduced Horoyah, an ongoing work on a knowledge graph-based framework designed to empower civil society and citizens in monitoring elections within the challenging environments of LMICs. The knowledge graph is available online on Zenodo[10]. We have argued that dysfunctions in electoral

[8] https://www.techpolicy.press/security-and-reliability-concerns-around-internet-voting-outweigh-benefits/.
[9] https://www.brennancenter.org/our-work/analysis-opinion/election-year-risks-ai.
[10] https://doi.org/10.5281/zenodo.16879062.

processes erode public trust and that ICT, when applied thoughtfully, can be a powerful tool for transparency and accountability. By providing a structured, auditable, and accessible platform for citizen-reported data, Horoyah enables CSOs and the public to play a more effective role in safeguarding democratic integrity.

The use case of the 2020 Guinea election illustrates the platform's potential to provide an independent, evidence-based counter-narrative to official results, thereby fostering public confidence and holding institutions accountable. However, we recognize that the path to effective implementation is filled with challenges, including the pervasive threats of disinformation, cybersecurity risks, and the digital divide. Overcoming these obstacles requires a collaborative, context-sensitive approach involving CSOs, technologists, media organizations, and electoral bodies.

As we look to the future, the rise of AI presents both new threats and new opportunities. The goal must be to steer the development of these powerful technologies towards strengthening democratic institutions. Frameworks like Horoyah represent a crucial step in this direction—harnessing the power of data and technology not to replace human judgment, but to empower citizen engagement and build a more transparent and trustworthy democratic future. That is why for low-literacy users, the future work of the Horoyah project envisions using AI-powered voice-based applications to make the reporting process more inclusive and invest in educating end-users to deal with the platform.

Acknowledgments. The author acknowledges the Union of Free Radios and Televisions of Guinea (URTELGUI) for providing the 2020 sample dataset, which was instrumental in illustrating the use case on the presidential election. Furthermore, the author extends gratitude to Dr. Rabia Azzi for his technical contributions to the development of the Horoyah dashboard. Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baader, F., Knechtel, M., Peñaloza, R.: A generic approach for large-scale ontological reasoning in the presence of access restrictions to the ontology's axioms. In: Hutchison, D., et al. (eds.) The Semantic Web - ISWC 2009, vol. 5823, pp. 49–64. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_4
2. Bouras, C., Katris, N., Triantafillou, V.: An electronic polling service to support public awareness using web technologies. In: ECIS 2001 Proceedings (2001). https://aisel.aisnet.org/ecis2001/14
3. Chondros, N., et al.: D-DEMOS: a distributed, end-to-end verifiable, internet voting system. In: 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), pp. 711–720. IEEE, Nara (2016). https://doi.org/10.1109/ICDCS.2016.56. http://ieeexplore.ieee.org/document/7536569/

4. Contributors, O.: Planet dump (2017). https://planet.osm.org. https://www.openstreetmap.org
5. Erdely, A.: A copula based approach for electoral quick counts (2019). https://doi.org/10.13140/RG.2.2.17665.71526. http://rgdoi.net/10.13140/RG.2.2.17665.71526
6. Hogan, A., et al.: Knowledge graphs. ACM Comput. Surv. **54**(4), 1–37 (2022). https://doi.org/10.1145/3447772. https://dl.acm.org/doi/10.1145/3447772
7. Kalaris, K., Wong, G., English, M.: Understanding networks in low-and middle-income countries' health systems: a scoping review. PLOS Glob. Public Health **3**(1), e0001387 (2023). https://doi.org/10.1371/journal.pgph.0001387. https://dx.plos.org/10.1371/journal.pgph.0001387
8. Ladner, A., Pianzola, J.: Do voting advice applications have an effect on electoral participation and voter turnout? Evidence from the 2007 Swiss federal elections. In: Tambouris, E., Macintosh, A., Glassey, O. (eds.) Electronic Participation, vol. 6229, pp. 211–224. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15158-3_18
9. Matharu, G.S., Mishra, A., Chaudhary, L.: Integrated election voting system: a model for leveraging ICT in the Indian election scenario. In: Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, pp. 1–7. ACM, Udaipur (2014). https://doi.org/10.1145/2677855.2677944. https://dl.acm.org/doi/10.1145/2677855.2677944
10. Ouedraogo, I., Some, B.M.J., Oyibo, K., Benedikter, R., Diallo, G.: Evaluating a phone-based Interactive Voice Response system for reducing misinformation and improving malaria literacy. In: Information Technology for Development, pp. 1–27 (2024). https://doi.org/10.1080/02681102.2024.2414193. https://www.tandfonline.com/doi/full/10.1080/02681102.2024.2414193
11. Rabbani, K., Lissandrini, M., Hose, K.: SHACL and ShEx in the wild: a community survey on validating shapes generation and adoption. In: Companion Proceedings of the Web Conference 2022, pp. 260–263. ACM, Virtual Event, Lyon (2022). https://doi.org/10.1145/3487553.3524253. https://dl.acm.org/doi/10.1145/3487553.3524253
12. Schryen, G.: E-Democracy: Internet Voting. Universität Regensburg (2003). https://doi.org/10.5283/EPUB.21239. https://epub.uni-regensburg.de/id/eprint/21239
13. Singer, S., Tsai, E.: electiondata: a Python package for consolidating, checking, analyzing, visualizing and exporting election results. J. Open Source Softw. **7**(69), 3739 (2022). https://doi.org/10.21105/joss.03739. https://joss.theoj.org/papers/10.21105/joss.03739

Semantic Similarity Analysis of Theses Using Sentence Embeddings and Knowledge Graphs

Ibeth J. del Pilar Cusi[1(✉)], Lizeth C. Mamani[1], Gabriela Zuñiga-Rojas[1,2], Edwin Alvarez-Mamani[1,3], and Harley Vera-Olivera[1]

[1] Universidad Nacional de San Antonio Abad del Cusco, Cusco, Peru
{210924,210934,gabriela.zuniga,edwin.alvarez,harley.vera}@unsaac.edu.pe, gzunigar@uandina.edu.pe
[2] Universidad Andina del Cusco, Cusco, Peru
[3] Universidad Peruana de Ciencias Aplicadas, Lima, Peru

Abstract. Theses authored by undergraduate students at Spanish-speaking universities are often omitted from major academic databases such as Scopus and Web of Science. These platforms prioritize English-language publications, including indexed journal articles and papers published in the proceedings of major conferences. As a result, these works, often written in Spanish and stored in institutional repositories, are underrepresented in scientific information systems, making the review of the state of the art a challenging task, particularly when aiming to include local and undergraduate academic contributions. This work presents a Knowledge Graph (KG) constructed from 7 739 thesis authored by undergraduate students at the Universidad Nacional de San Antonio Abad del Cusco (UNSAAC), Perú. The documents, written in Spanish, were represented using vector embeddings generated by pretrained language models. Semantic similarity between documents was computed using cosine distance. Based on these similarities, a KG was built and modeled in `Neo4j`. The KG was further enriched with institutional metadata, including faculty and academic programs. This structure allows for the identification of conceptual links among documents and the detection of thematic communities within the collection. The source code, datasets, and implementation details are available at: https://github.com/JanelaCusi-03/semantic-thesis-similarity-KG.

Keywords: knowledge graphs · sentence embeddings · semantic similarity · theses

1 Introduction

A thorough review of the state of the art is a critical step in the development of any research project. It should be conducted systematically at local, national,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 93–108, 2026.
https://doi.org/10.1007/978-3-032-20322-9_7

and international levels. This process contextualizes the research problem and identifies knowledge gaps that the proposed study aims to address [3,16]. International literature reviews commonly rely on academic databases such as Scopus, Web of Science, Redalyc, and Dialnet. These platforms prioritize English-language publications, including indexed journal articles and papers published in the proceedings of major conferences. As a result, they often exclude valuable academic outputs such as thesis, which are typically stored in institutional repositories. Although essential for accessing local and national research, these repositories present technical limitations, including exact-match search, poor handling of synonyms and spelling variants, and inconsistent metadata structures. These issues hinder efficient retrieval and complicate the identification of original contributions. Initial efforts have been proposed to organize research projects through structured metadata and classification systems [15,17]. Other approaches have explored the use of KG to represent and link scientific contributions [9,13,14]. However, these initiatives have not yet integrated theses stored in institutional repositories, particularly those written in Spanish, despite their potential to enrich academic knowledge bases at the undergraduate and graduate levels.

This work presents a KG constructed from 7 739 theses authored by undergraduate students and stored in the institutional repository of the UNSAAC. The KG represents semantic similarity between documents written in Spanish, using vector representations generated by pretrained language models and cosine distance as the similarity metric. The resulting semantic graph was modeled in `Neo4j` and enriched with institutional metadata such as faculty and academic school. Its structure enables the detection of thematic communities and supports the implementation of a semantic search system. Additionally, the KG could support trend analysis, identification of topic gaps, recommendation of related works, and mapping of collaborations or affinities between students and advisors.

The remainder of this article is organized as follows: Sect. 2 reviews related works on academic knowledge graphs and sentence embeddings, highlighting the current limitations in processing Spanish-language academic content. Section 3 details the methodology, describing each phase of the proposed pipeline, from data extraction and preprocessing to embedding generation, dimensionality reduction, and knowledge graph construction. Section 4 presents the experimental results, including quantitative evaluations and visual analyses of model performance, followed by a description of the final graph structure. Finally, Sect. 5 discusses the conclusions drawn from this study and outlines potential applications and future research directions.

2 Related Works

Recent advances in KG applications for academic information management have demonstrated their effectiveness in organizing and analyzing scholarly content. The development of large-scale academic graphs such as Microsoft Academic

Graph (MAG) [15] and Semantic Scholar Academic Graph (S2AG) [14] has established robust frameworks for representing scholarly entities and their relationships. These systems build upon semantic ranking techniques [17] that leverage knowledge graph embeddings to enhance search capabilities beyond traditional word-based methods. A significant innovation in this domain involves the integration of textual corpora with structured knowledge graphs [1], where text provides broad contextual information while KGs offer precise relational data, creating complementary representations that improve semantic understanding. Further developments have extended these approaches to specialized academic domains. Educational knowledge graphs [7] have adapted embedding techniques to represent learning resources and curricula, while scientific trend analysis systems [8] employ transformer models to extract and visualize emerging research topics. The ongoing evolution of open knowledge graphs like OpenAlex [9] continues to refine author disambiguation and citation network accuracy, addressing critical challenges in scholarly data representation.

While these contributions have advanced academic knowledge management, important limitations persist regarding language support, computational efficiency, and institutional context integration. To address these limitations, our work introduces three key innovations: 1) the first systematic evaluation of Spanish-language sentence embeddings for academic content, filling a critical gap in multilingual KG applications; 2) an autoencoder-based dimensionality reduction method that reduces embedding sizes, enabling deployment in resource-constrained institutional repositories; and 3) a hierarchical knowledge graph architecture that explicitly models relationships between documents (theses), academic programs, and faculties, capturing both semantic similarities and institutional contexts. This integrated approach advances beyond current systems by providing a scalable framework specifically optimized for local academic ecosystems where semantic precision and computational efficiency are paramount.

3 Methodology

Figure 1 provides a graphical summary of the proposed methodology, which follows a sequential workflow encompassing automated thesis collection through knowledge graph construction. The process comprises the following stages: data extraction from the institutional repository; corpus preprocessing; embedding generation using six distinct language models; dimensional reduction of embeddings; a comparative performance analysis across models, and semantic graph generation within a graph-oriented database.

3.1 Corpus Extraction and Construction

Since the institutional repository of UNSAAC[1] does not provide a public API for bulk record access, an automated data collection process was imple-

[1] https://repositorio.unsaac.edu.pe/.

mented using web scraping techniques. A `Python` script was developed employing the `BeautifulSoup` library, which iteratively accessed repository pages and extracted metadata from each thesis.

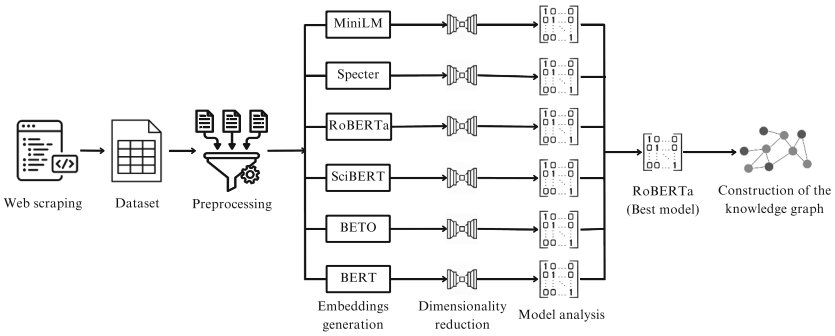


Fig. 1. Semantic pipeline for similarity analysis in thesis.

Once collected, the data were organized into a tabular structure (see Table 1), with each row representing a thesis and each column corresponding to a specific attribute. This organization facilitated both manual inspection and processing using `Python` analysis tools. The initial corpus consisted of 9 580 theses written in Spanish, extracted from the UNSAAC institutional repository.

Table 1. Structure of extracted metadata per thesis.

Field	Description
Id	Unique thesis identifier
Title	Complete project title
Abstract	Brief description of content and objectives
Keywords	Representative content terms
Publication year	Year of project defense
Academic program	Degree-granting academic program
Faculty	Corresponding UNSAAC faculty
URL	Online access link

3.2 Dataset Preprocessing

The dataset underwent a cleaning and refinement process to ensure its quality and semantic relevance. The following procedures were applied: duplicate

removal, irrelevant content filtering, and validation of essential fields. In duplicated removal, duplicate records were removed based on the *title* field, ensuring the uniqueness of each thesis. Subsequently, irrelevant content filtering, records corresponding to documents other than theses were discarded. Finally, in validation of essential fields, each thesis was verified to contain a *title*, *abstract*, and at least one *keyword*, minimum requirements for subsequent semantic representation. As a result, a curated corpus of 7 739 theses was obtained, all consistent and suitable for semantic vectorization in subsequent stages of the proposed model.

3.3 Embedding Generation Using Language Models

To capture precise semantic representations of the theses, six sentence embedding models were employed, each based on different architectures and training sets tailored to specific domains or languages. All models were applied to the same curated dataset, consisting of 7 739 theses written in Spanish. The input provided to each model consisted of the concatenation of the *title*, *abstract*, and *keywords* of each thesis. This strategy enabled a more comprehensive capture of the thematic and linguistic information contained in the documents. The models used were as follows:

- all-MiniLM-L6-v2-similarity-es (`MiniLM_ES`), a Spanish-adapted version of `MiniLM` optimized for semantic similarity tasks with low computational cost [10].
- `SPECTER`, trained on titles and abstracts of scientific articles, designed for academic document retrieval and comparison [5].
- scibert-nli (`SciBERT_NLI`), a `SciBERT` variant oriented toward semantic inference tasks in scientific texts, enhancing the detection of relationships between technical documents [2].
- roberta-base-bne (`RoBERTa_BNE`), a `RoBERTa` model trained on a Spanish corpus derived from institutional and scientific sources, suitable for natural language processing tasks in Spanish [6].
- bert-base-spanish-wwm-cased (`BETO`), a BERT-based model trained from scratch on Spanish data using Whole Word Masking, which improves full-context capture [4].
- bertin-roberta-base-spanish (`BERTIN`), a `RoBERTa` variant trained on extensive corpora from European and Latin American Spanish, optimized for in-depth language analysis [11].

Each model generated a high-dimensional vector matrix, where each vector encodes the semantic information of an individual thesis.

3.4 Dimensionality Reduction Using Autoencoders

Since the evaluated language models generate embeddings with different dimensions (384 and 768), dimensional reduction was implemented to unify their representation into a common latent space of 256 dimensions. This approach enables

fair comparison between models and mitigates potential biases arising from structural differences, while optimizing computational efficiency for subsequent stages such as similarity matrix calculation and graph analysis over the extensive corpus ($N \approx 7\,739$).

An independent dense autoencoder was trained for each embedding set, selecting this architecture for its capacity to learn nonlinear compressions that preserve essential semantic structure. Each autoencoder implements a symmetric, fully connected encoder-decoder architecture, utilizing ReLU activation functions in hidden layers and linear output in the decoding stage. Optimization was performed using the Adam algorithm with a learning rate of 1×10^{-3}, incorporating L2 regularization and dropout ($p = 0.2$) to enhance generalization. Training extended for a maximum of 100 epochs with early stopping (patience of 10 epochs), monitoring validation loss. Data was split into 90% for training and 10% for validation, preserving the proportion of theses per faculty.

The quality of the reduction was evaluated through: i) reconstruction loss (MSE), which quantifies the ability of the model to reproduce the original embeddings, and ii) Spearman correlation between original and reduced similarity matrices. Results showed high correlation ($\rho > 0.8$ for `RoBERTa_BNE`, `SciBERT_NLI` and `SPECTER` models), confirming that the reduction consistently preserves semantic relationships between theses.

3.5 Semantic Similarity Computation

After obtaining the reduced vector representations from sentence embedding models, we computed semantic similarity across all theses in the corpus. The vector representations, stored in *.parquet* format, were processed using cosine similarity; this metric measures semantic proximity between vector pairs regardless of their magnitude [12].

The formula used for cosine similarity between two vectors A and B is defined in Eq. (1):

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{1}$$

For each of the six evaluated models, we generated a $7\,739 \times 7\,739$ similarity matrix where each cell (i, j) represents the semantic similarity degree between theses i and j. These matrices were stored in *.parquet* format, enabling efficient in-memory processing and facilitating subsequent comparative analysis and visualization.

3.6 Comparative Analysis of Embedding Models

The similarity matrices generated by each model were quantitatively evaluated using five key metrics designed to assess the quality of thematic clustering and semantic fidelity within the corpus of theses:

- Silhouette index: Simultaneously evaluates intra-cluster cohesion and inter-cluster separation. Higher values indicate more distinct and well-separated groupings.
- Similarity quantile distribution: Enables observation of the density and dispersion of cosine similarity values produced by each model, reflecting their ability to distinguish between theses with varying levels of thematic affinity.
- Spearman matrix correlation: Measures the relative consistency between similarity patterns produced by different models, indicating their agreement in ranking thesis pairs by semantic similarity.
- Clustering quality: Assesses the structural organization of the formed semantic groups. This metric is based on two components: *intra-cluster distance*, which measures the internal compactness of clusters (i.e., how closely grouped the theses are within each cluster), and *inter-cluster distance*, which measures the separation between clusters (i.e., how distinct the clusters are from one another).
- Davies-Bouldin index (DBI): Computes a ratio between intra-cluster dispersion and inter-cluster separation. Lower DBI values indicate superior clustering quality, representing an optimal balance between compactness and separation.

The objective of this analysis was to determine which of the sentence embedding models most accurately captured the semantic relationships present within the set of theses.

3.7 Knowledge Graph Construction in `Neo4j`

The next step involved constructing a knowledge graph from the generated similarity matrices. Using the graph-oriented database `Neo4j`, theses were modeled as nodes, and their semantic similarity relationships were represented as edges labeled *SIMILAR*.

These edges were created only for document pairs whose cosine similarity exceeded a threshold of 0.95, allowing the graph to capture thematically meaningful connections without overloading the network with irrelevant links. This threshold is supported by previous studies recommending high values (between 0.8 and 0.95) for constructing dense and relevant semantic graphs in academic contexts [18].

A set of `Python` scripts was used for data loading, using the `py2neo` library, which enables programmatic interaction with the `Neo4j` database through the `Cypher` query language. This tool facilitated the automated creation of nodes and relationships, allowing the semantic structure extracted from the similarity matrices to be replicated within the graph. In addition to theses, supplementary nodes representing relevant organizational entities within UNSAAC were incorporated. Specifically, nodes were added for each academic program and faculty, establishing *BELONGS_ TO* relationships between theses and their corresponding academic units. This enabled cross-analysis between thematic content and institutional structure. Figure 2 presents a conceptual diagram of the core structure of the KG, illustrating the types of nodes and relationships employed.

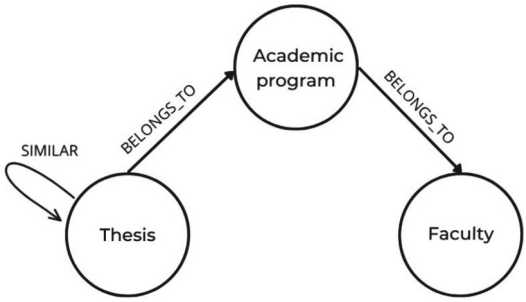


Fig. 2. Core structure of the knowledge graph built in `Neo4j`.

4 Results

The experiments were conducted in `Google Colaboratory`, using `Python 3.10` and specialized libraries such as `Transformers`, `Sentence-Transformers`, `Pandas`, `UMAP-learn`, `Scikit-learn`, `Seaborn`, `Matplotlib`, and `Py2Neo`. This environment enabled the execution of semantic vectorization, dimensionality reduction, comparative evaluation, and results visualization tasks. For the construction of the knowledge graph, `Neo4j Aura`, a cloud-based graph-oriented database, was utilized. The graphs were generated from *.csv* files processed in Colab through `Py2Neo` scripts and `Cypher` queries, facilitating an efficient and reproducible implementation of the complete semantic analysis workflow on the corpus of 7 739 theses.

4.1 Quantitative Evaluation of Semantic Similarity Models

To select the most suitable model for capturing similarity between Spanish-language theses, the performance of six models was evaluated: `BETO`, `SPECTER`, `BERTIN`, `MiniLM_ES`, `RoBERTa_BNE`, and `SciBERT_NLI`, using clustering metrics, similarity distributions, and correlation between representations.

The original results (see Table 2) highlight `RoBERTa_BNE` as the superior model. Its high Silhouette Index (0.5507) and low Davies-Bouldin Index (DBI) (0.5912) indicate that the generated vectors form cohesive and clearly separated internal clusters, favoring acceptable thematic segmentation. Furthermore, the highest intercluster distance (8.1220) reinforces its capacity to clearly differentiate between distinct topics, while the relatively high median similarity (0.8642) demonstrates that it maintains coherence within clusters without overgeneralizing.

In contrast, `BERTIN` exhibits the highest median similarity (0.9992), evidencing that nearly all documents are considered extremely similar. This tendency toward overgeneralization reduces its capacity to discriminate between distinct content, limiting its utility for detailed semantic exploration tasks. `MiniLM_ES` and `SciBERT_NLI` show adequate similarity dispersion, but their clustering met-

rics, particularly lower intercluster distances and higher DBI values, indicate reduced capacity to organize the corpus thematically in a consistent manner.

BETO and SPECTER present intermediate performance: they generate reasonably cohesive clusters, but their intercluster distances and DBI suggest that the separation between topics is not as pronounced as in RoBERTa_BNE. Spearman correlation supports these findings, evidencing that RoBERTa_BNE provides a more stable and coherent semantic representation.

When considering the normalized and consolidated results in Table 3, RoBERTa_BNE obtains the highest composite score, reaffirming its leadership as the optimal model for representing semantic relationships in theses. The remaining models, while useful in specific contexts, present limitations: BERTIN tends to overgeneralize, MiniLM_ES and SciBERT_NLI lack sufficient thematic separation, and BETO and SPECTER offer reasonable but inferior performance.

Table 2. Quantitative evaluation of semantic similarity models.

Model	Silhouette ↑	Similarity ↑	Spearman ↑	Intra ↓	Inter ↑	DBI ↓
RoBERTa_BNE	**0.5507**	**0.8642**	**0.5267**	1.5509	**8.1220**	**0.5912**
BETO	0.5273	0.8207	**0.4827**	1.6984	6.9780	0.7196
SPECTER	0.5032	0.8426	**0.4266**	1.7650	6.3307	0.6752
SciBERT_NLI	0.4839	0.8400	0.3591	1.4876	4.9058	0.8079
MiniLM_ES	0.4283	0.8525	0.2089	**1.0863**	4.7208	0.7286
BERTIN	0.4086	**0.9992***	0.2682	**0.9491**	3.5295	0.8240

Global Ranking, to integrate multiple metrics into a coherent comparative evaluation, we applied a normalization process to all variables with equal weighting assigned to each metric. Table 3 presents the consolidated results of the global ranking analysis. In this comprehensive evaluation, the RoBERTa_BNE model achieved the highest composite score, emerging as the most robust alternative for capturing and representing semantic relationships among theses in the analyzed corpus.

4.2 Visual Analysis of Similarity and Thematic Clustering

To complement the quantitative evaluation, we generated visualizations to examine the semantic behavior of each model in representing theses. This analysis employed three visual approaches: similarity histograms, heatmaps, and two-dimensional Uniform Manifold Approximation and Projection (UMAP).

Similarity Distribution, a composite histogram figure was created for all six models (see Fig. 3), using a sample of up to 10 million similarity pairs (excluding diagonal elements). BERTIN exhibited a strongly right-skewed distribution

Table 3. Model ranking based on normalized metrics.

Model	Silhouette↑	Similarity↑	Spearman↑	Intra↓	Inter↑↑	DBI↓	Score
`RoBERTa_BNE`	1.0000	0.2437	1.0000	0.2624	1.0000	1.0000	**0.6437**
`BETO`	0.8353	0.0000	0.8615	0.0816	0.7509	0.4485	0.4254
`SPECTER`	0.6657	0.1227	0.6850	0.0000	0.6100	0.6392	0.3889
`BERTIN`	0.0000	1.0000	0.1866	1.0000	0.0000	0.0000	0.3124
`SciBERT_NLI`	0.5299	0.1081	0.4726	0.3400	0.2997	0.0692	0.2599
`MiniLM_ES`	0.1386	0.1782	0.0000	0.8318	0.2594	0.4098	0.2597

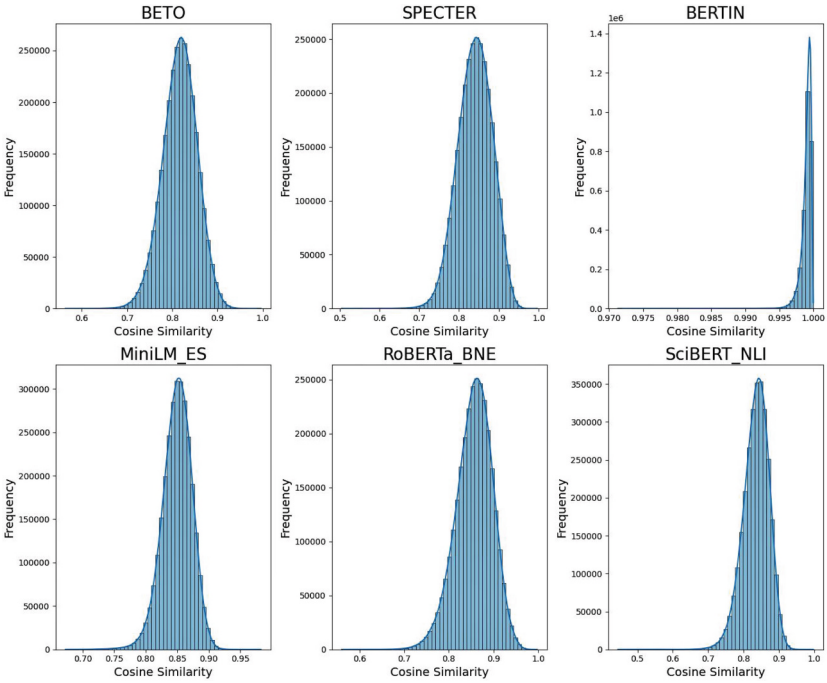


Fig. 3. Cosine similarity distribution for each model.

concentrated near 1.0, indicating limited semantic discrimination. In contrast, `RoBERTa_BNE`, `SPECTER`, and `BETO` showed more balanced distributions, suggesting superior capability to identify genuine semantic similarities among theses.

Heatmaps, Fig. 4 displays 1000×1000 submatrix heatmaps for each model. Visualizations reveal global clustering patterns, with RoBERTa_BNE, BETO, and `SPECTER` showing greater structural variability - indicative of better semantic differentiation. `BERTIN` produced near-homogeneous matrices, confirming its limited separation capability.

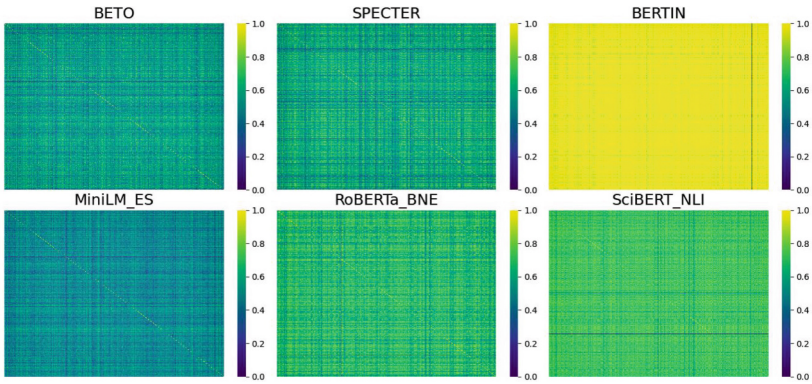


Fig. 4. Similarity heatmaps for each model.

Faculty-based UMAP Visualization, UMAP dimensionality reduction was applied to similarity matrices, with points colored by Faculty affiliation (Fig. 5). `RoBERTa_BNE`, `BETO`, and `SPECTER` achieved clearer cluster segmentation, demonstrating semantic representations aligned with institutional divisions. `BERTIN` and MiniLM_ES showed point overlapping, suggesting unintended conceptual blending.

Academic program-based UMAP visualization, the analysis was repeated using academic program as the coloring scheme (see Fig. 6). While group separation decreased due to more specific categories, `RoBERTa_BNE` maintained superior theses organization, confirming its capacity to capture topics with high semantic precision (see the interactive visualizations in the link provided in the footnote)[2].

4.3 Knowledge Graph Visualization

Based on the obtained results, the `RoBERTa_BNE` model was selected to construct an *academic semantic knowledge graph* based on a corpus of 7 739 theses. The structure was implemented in the `Neo4j AuraDB` database, enabling explicit representation of semantic similarity relationships through a network of nodes and edges.

The graph comprises three node types: i) theses with 7 739 nodes, ii) professional schools with 40 nodes, and iii) faculties with 10 nodes. Relations (181 053 in total) were modeled as: *SIMILAR* (173 274 in total), connecting pairs of theses with cosine similarity exceeding 0.95, and *BELONGS_TO* (7 779 in total), linking each thesis to its respective academic unit.

[2] Interactive versions of the UMAP projections available at: https://github.com/JanelaCusi-03/semantic-thesis-similarity-KG.

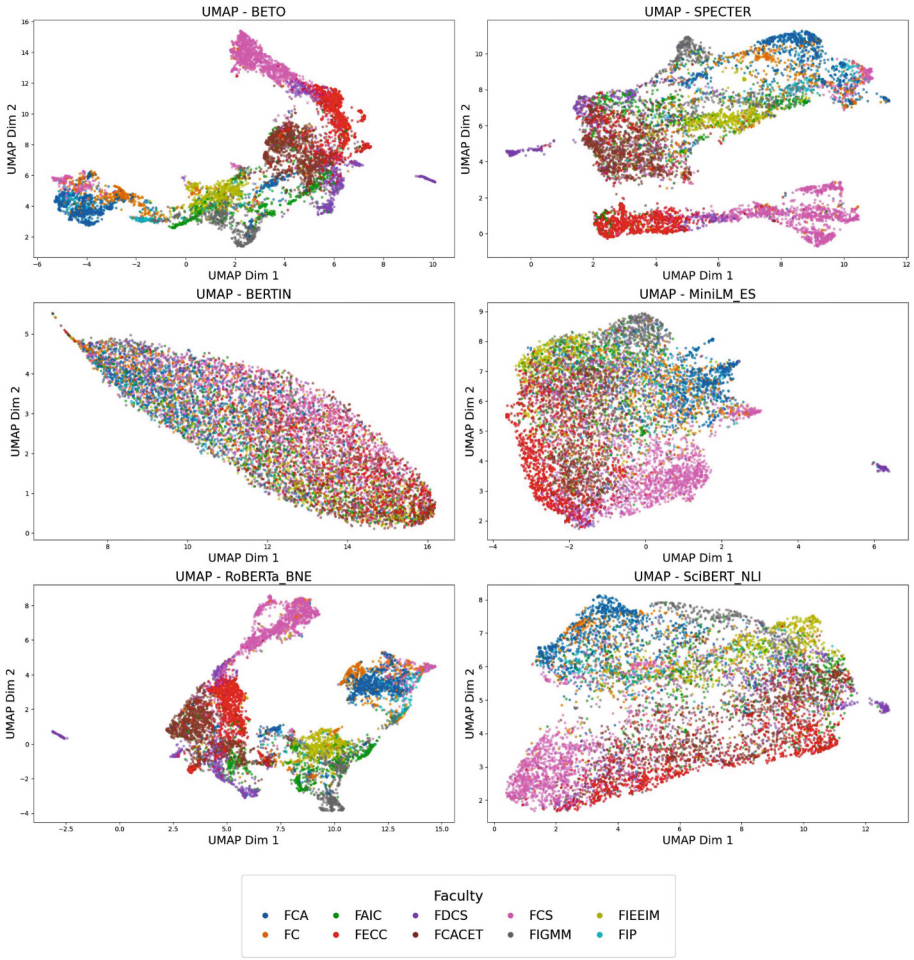


Fig. 5. Dimensionality reduction by faculty. Note: FCA - Facultad de Ciencias Agrarias, FC - Facultad de Ciencias, FAIC - Facultad de Arquitectura e Ingeniería Civil, FECC - Facultad de Educación y Ciencias de la Comunicación, FDCS - Facultad de Derecho y Ciencias Sociales, FCACET - Facultad de Ciencias Administrativas, Contables, Económicas y Turismo, FCS - Facultad de Ciencias de la Salud, FIGMM - Facultad de Ingeniería Geológica, Minas y Metalúrgica, FIEEIM - Facultad de Ingeniería Eléctrica, Electrónica, Informática y Mecánica, FIP - Facultad de Ingeniería de Procesos.

Visual Exploration. Figure 7 presents a partial view of the graph, generated from thematic subgraphs extracted from Neo4j. Colors indicate entity type (thesis, school, or faculty), while connections reflect relevant semantic relationships among theses. This visualization facilitates interpretation of the corpus's thematic structure and identification of regions with high interconnectivity.

Structure and Density. The resulting graph contains a total of 173 274 *SIMILAR* edges, evidencing high semantic connectivity among theses. This density enabled

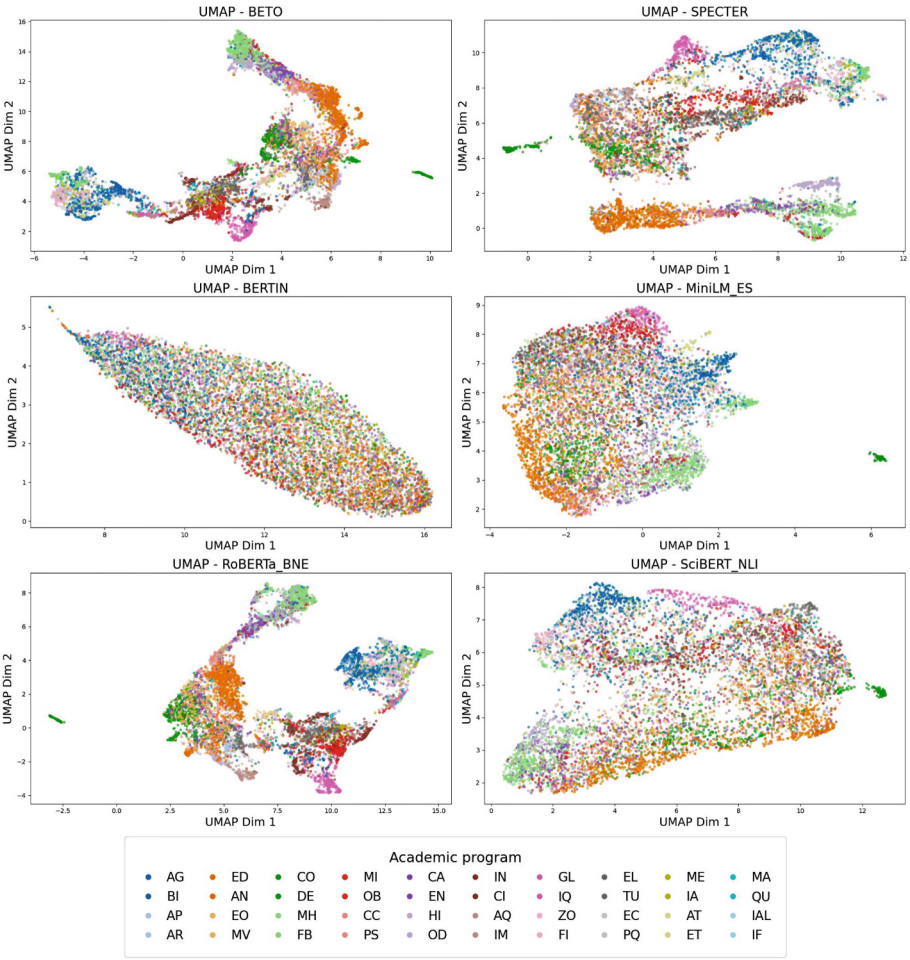


Fig. 6. `UMAP` visualization of theses grouped by academic program. AG - Agronomía, BI - Biología, AP - Ingeniería Agropecuaria, AR - Arquitectura, ED - Educación Cusco, Canas y Espinar -, AN - Antropología, EO - Economía, MV - Medicina Veterinaria, CO - Contabilidad, DE - Derecho, MH - Medicina Humana, FB - Farmacia y Bioquímica, MI - Ingeniería de Minas, OB - Obstetricia Andahuaylas -, CC - Ciencias de la Comunicación, PS - Psicología, CA - Ciencias Administrativas, EN - Enfermería, HI - Historia, OD - Odontología, IN - Ingeniería Informática y de Sistemas, CI - Ingeniería Civil, AQ - Arqueología, IM - Ingeniería Metalúrgica, GL - Ingeniería Geológica, IQ - Ingeniería Química, ZO - Zootecnia, FI - Física, EL - Ingeniería Eléctrica, TU - Turismo, EC - Ingeniería Electrónica, PQ - Ingeniería de Petroquímica, ME - Ingeniería Mecánica, IA - Ingeniería Agroindustrial, AT - Agronomía Tropical, ET - Ecoturismo Quillabamba -, MA - Matemática, QU - Química, IAL - Ingeniería en Industrias Alimentarias, IF - Ingeniería Forestal.

identification of regions with strong internal cohesion, associated with natural thematic clusters. Unlike traditional clustering methods, these groupings emerge organically from calculated similarity relationships, without requiring explicit partitioning algorithms.

Identification of Thematic Communities. Community detection was performed through semantic queries on the graph, exploring subsets of highly connected nodes. Significant groupings were observed around disciplinary areas such as *Engineering*, *Accounting Sciences*, *Health*, and *Social Sciences*. These results demonstrate the graph's capacity to reflect the thematic structure of the institutional corpus and facilitate retrieval of interdisciplinary knowledge.

For instance, queries executed in `Neo4j` enabled identification of *Faculty of Law* projects related to *Social Sciences* or *Economics*, evidencing the graph's potential to uncover non-trivial semantic connections.

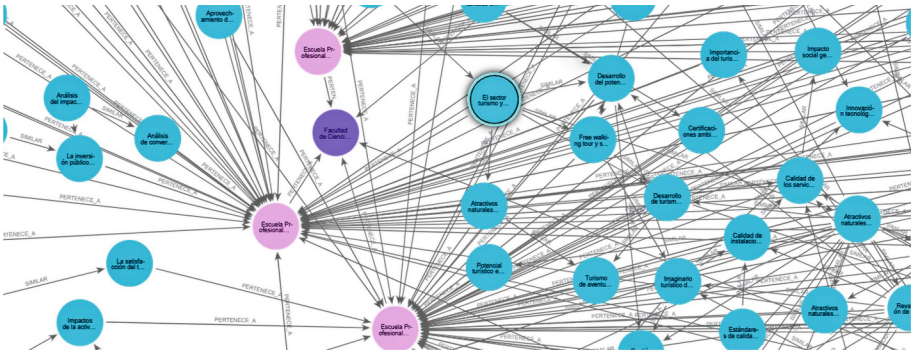


Fig. 7. Partial view of the knowledge network generated with `RoBERTa_BNE`.

5 Conclusions

This study presented a comprehensive evaluation of six semantic representation models applied to a corpus of 173 274 theses written in Spanish. Based on quantitative metrics, `RoBERTa_BNE` was identified as the model that most effectively captures the thematic structure of the repository and semantic relationships among theses.

As a practical application, an academic knowledge graph was constructed from the semantic vectors generated by the `RoBERTa_BNE` model. This graph integrates nodes corresponding to theses, schools, and faculties, connected through semantic similarity and institutional affiliation relations. The resulting representation enables analysis of thematic patterns, discovery of hidden relationships among theses, and constitutes a solid foundation for advanced tools that overcome the limitations of traditional search systems.

In conclusion, this work demonstrates how the combination of advanced language models and knowledge graphs can transform the management, analysis, and discovery of academic information. The methodology and results presented here establish a robust framework that can be replicated and extended, paving

the way toward integration of semantic intelligence in large-scale academic repositories. To enhance and advance this work, we consider the following future directions:

- Comprehensive graph evaluation: Conduct an integral evaluation of the generated knowledge graph through expert domain validation. A representative sample of entities and relations will be assessed to determine semantic accuracy, thematic coverage, and practical utility for institutional knowledge exploration.
- Graph enrichment: Incorporate new entities, such as *Author* and *Advisor*, to model and analyze collaboration and co-authorship networks, enabling a deeper understanding of institutional research dynamics.
- Interinstitutional expansion: Apply the methodology to repositories from other universities, aiming to construct a federated academic knowledge graph that provides a more comprehensive view of research landscape at the regional level.
- Real-time system implementation: Develop an automated pipeline that continuously updates the graph with new theses as they are published, ensuring that the resource remains current and useful for students, faculty, and researchers.
- Development of value-added applications: Leverage the knowledge graph as the core infrastructure for end-user oriented tools, including:
  - Intelligent recommendation systems: Automate the suggestion of related theses and projects based on semantic similarity relationships, facilitating exploration of relevant bibliography.
  - Academic chatbot: Integrate a conversational assistant capable of answering natural language queries about theses, authors, topics, and their interrelations, democratizing access to academic information.

References

1. Allen, C.S.: Towards a Theoretical Understanding of Word and Relation Representation. Ph.D. thesis (2021)
2. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), p. 3613. Association for Computational Linguistics (2019)
3. Berndtsson, M., Hansson, J., Olsson, B., Lundell, B.: Thesis projects: a guide for students in computer science and information systems. Springer (2008)
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pretrained bert model and evaluation data. arXiv preprint arXiv:2308.02976 (2023)
5. Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.: Specter: document-level representation learning using citation-informed transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2020)

6. Gutiérrez Fandiño, A., et al.: Maria: Spanish language models. Procesamiento del Lenguaje Nat. **68** (2022)
7. Karray, M.H., et al.: Automatic construction of educational knowledge graphs: a word embedding-based approach. Information **14**, 526 (2023). https://doi.org/10.3390/INFO14100526, https://www.mdpi.com/2078-2489/14/10/526/htm
8. Muñoz, R., Gutiérrez, Y., Montoyo, A.: T2know: an advance scientific-tecnical text analysis platform for trend and knowledge extraction using nlp techniques (2023). http://ceur-ws.org
9. Priem, J., Piwowar, H., Orr, R.: Openalex: a fully-open index of scholarly works, authors, venues, institutions, and concepts (2022). http://arxiv.org/abs/2205.01833
10. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), p. 3982. Association for Computational Linguistics (2019)
11. De la Rosa, J., Ponferrada, E.G., Romero, M., Villegas, P., de Prado Salas, P.G., Grandury, M.: Bertin: Efficient pre-training of a spanish language model using perplexity sampling. Procesamiento del Lenguaje Natural **68**, 13–23 (2022)
12. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge (2008)
13. Tosi, M.D.L., Reis, J.C.D.: Scikgraph: A knowledge graph approach to structure a scientific field. J. Inf. **15**, 101109 (2021). https://doi.org/10.1016/J.JOI.2020.101109, https://www.sciencedirect.com/science/article/abs/pii/S175115772030626X
14. Wade, A.D.: The semantic scholar academic graph (s2ag) pp. 739–739 (2022). https://doi.org/10.1145/3487553.3527147
15. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: when experts are not enough. Quant. Sci. Stud. **1**, 396–413 (2020). https://doi.org/10.1162/QSS_A_00021
16. Wazlawick, R.S.: Metodologia de pesquisa para ciência da computação, vol. 3ra. Edi, Elsevier Rio de Janeiro (2020)
17. Xiong, C., Power, R., Callan, J.: Explicit semantic ranking for academic search via knowledge graph embedding (2017). https://doi.org/10.1145/3038912.3052558
18. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning based recommender system: a survey and new perspectives. ACM Comput. Surv. (CSUR) **52**(1), 1–38 (2019)

Measuring Healthcare Accessibility Through Clustering Techniques and Multi-layer Network Analysis

Dhananjay Balakrishnan[1], Svea Drekshagen[1], Vedant Sahu[1], Mersedeh Tariverdi[2], and Miguel Nunez-del-Prado[2,3(✉)]

[1] Institute of Computational and Mathematical Engineering, Stanford University, California, USA
[2] The World Bank, Washington D.C., USA
m.nunezdelpradoco@worldbank.org
[3] Universidad Peruanas de Ciencias Aplicadas, Lima, Peru

Abstract. Accessibility to health services, both in-person and virtual, is vital for advancing Health and Well-being SDG. While e-healthcare can bridge some of these gaps, its success depends on reliable digital infrastructure and coverage. This paper provides an integrated approach that considers both dimensions, physical and digital. We evaluate Multi-Edge Graph model, Multi-Layer Graph model, Composite index, and unsupervised learning approaches to measure physical and digital accessibility apply for South Africa. Our findings show that clustering methods, reinforced by robust imputation, afford the best balance between objectivity, interpretability, and actionable insights under the current data environment offering a scalable framework for assessing hybrid healthcare systems and provide valuable insights for policy, planning, and resource allocation.

Keywords: Healthcare Accessibility · Clustering · Multilayer Graph

1 Introduction

Accessibility to health services, both in-person and virtual, is vital for advancing Health and Well-being Sustainable Development Goal. Healthcare accessibility is a multifaceted concept, encompassing an individual's capacity to obtain essential health services for prevention, diagnosis, treatment, and management of diseases [3].

Prior research, on physical accessibility, tries to further understand the complex relationship between distance and/or travel time in urban, rural and remote residents' access to healthcare services in developed countries [9]. Authors found that increased distance or travel time negatively impacted healthcare access; distance decay was observed at distances as short as 16.1 km or 30 min of travel time. Rural and remote residents were found to travel further and longer to access specialist health services compared to generalist service; distance decay

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 109–124, 2026.
https://doi.org/10.1007/978-3-032-20322-9_8

was observed more frequently in studies of generalist services (88%) than specialist services (80.6%). Beyond distance and travel time themselves, variables like age, Socio-economic status, gender, geographical barriers, ethnicity, and transportation variables were found to influence healthcare access. Chowdhury and Ravi [3] examine healthcare accessibility disparities in rural areas of developing countries, framing this as a critical global health challenge. The study's primary objective is to identify and analyze multidimensional barriers—including geographical, financial, and workforce-related constraints—that limit rural populations' access to essential health services. The authors synthesized secondary data from international reports and peer-reviewed literature. Key findings specific to time and distance barriers reveal that geographic isolation significantly impedes care: travel time/distance inversely correlates with health facility utilization, with inadequate transport infrastructure causing delayed emergency obstetric care, missed appointments, and poor chronic disease management. Notably, over 56% of rural residents globally lack timely access due to these spatial constraints, exacerbating preventable mortality. The study underscores that reducing travel burdens through infrastructural investments and telemedicine expansion is vital for equitable healthcare delivery.

Coombs et al. [4] explored rural healthcare providers' perspectives on social, cultural, and programmatic barriers to healthcare access in Montana, USA, with particular attention to the acceptability and appropriateness dimensions of access as outlined in Levesque's framework [6]. The study employed semi-structured interviews with a convenience sample of 12 healthcare providers in rural Montana. Data were collected between January and September 2021 via Zoom or phone, with interviews averaging 50 min in duration. Content analysis identified several key themes related to temporal and spatial barriers. Providers reported substantial challenges regarding accessibility linked to time and distance, noting that patients often traveled up to five hours, sometimes across state or county lines, to reach healthcare services, underscoring the geographical barriers characteristic of rural settings. Additionally, severe time limitations arose due to workforce shortages and excessive administrative burdens, such as navigating fragmented electronic medical records; only four providers reported being able to extend patient visit lengths beyond the standard duration, which was considered a "luxury". Therefore, the work of Tariverdi *et al.* [13] models individual's preferences, such as number of staffed beds, nurses, physicists, distance, etc. captured by a Huff model to account for the attractiveness of the facilities. Additionally road network infrastructure is incorporated for accessibility and road criticality measures.

With the expanding availability of internet connectivity, numerous studies have explored the impact of e-Health and telemedicine technologies on improving healthcare accessibility by overcoming barriers such as geographical distance, cost, and resource limitations. Judijanto *et al.* [12] assess how digital transformation enhances equitable access, especially for remote and underserved populations, finding that digital tools considerably reduce time- and distance-related barriers, although precise quantitative data remain limited. Similarly,

Anawade, Sharma, and Gahane [1] highlight that telemedicine eliminates travel burdens through virtual consultations and 24/7 availability, optimizing appointment scheduling and expanding access beyond conventional hours. Their review notes significant uptake, citing U.S. data where 57% of adults used telemedicine in 2021, with higher usage among rural communities. Further, Omaghomi *et al.* [10] focus on telemedicine in rural Africa, demonstrating that teleconsultations and remote diagnostics reduce the need for long-distance travel and associated economic burdens. However, infrastructural challenges—such as limited broadband, mobile device penetration, and inconsistent policy support—continue to constrain universal access. Collectively, these findings suggest that while e-Health integration holds substantial promise for enhancing healthcare accessibility, its success depends on coordinated infrastructure development, policy frameworks, and cross-sector collaboration.

Thus, few works propose a technical assessment of the digital infrastructure to evaluate the quality of service requirements of the e-health services. For example, Shao and Luo [5] developed the Enhanced Two-Step Virtual Catchment Area (E2SVCA) model to improve telehealth accessibility assessment by incorporating broadband speed requirements for video consultations, addressing limitations of the previous binary 2SVCA approach. Using data from Cook County, Illinois—including healthcare provider locations, population, broadband speeds, and road networks—they applied a weighted function based on Zoom's bandwidth thresholds. Their findings revealed that the original 2SVCA model overestimates accessibility, especially in underserved southern and northwestern areas with poor broadband quality, while central Chicago shows better access. The E2SVCA model emphasizes that broadband quality is as crucial as physical proximity for telehealth access, offering policymakers a refined tool to guide infrastructure investments and reduce healthcare digital inequities.

Similarly, Nunez-del-Prado et al. [11] proposed a spatiotemporal machine learning framework to assess e-health readiness across Peru, Sudan, and South Africa by integrating digital infrastructure quality with socio-economic indicators. Their analysis classified regions based on capacity to support digital health services, revealing significant disparities: urban centers showed strong connectivity suitable for advanced e-health applications, while rural and conflict-affected areas exhibited poor digital access, mostly limited to low-bandwidth services. Wealthier regions consistently had better network quality, though mobile networks were prone to greater temporal instability. This framework quantifies digital infrastructure gaps as key barriers to healthcare access and identifies priority zones for targeted investment to promote equitable e-health delivery. Together, these studies underscore the need to incorporate granular digital connectivity metrics and socio-economic context in evaluating and enhancing telehealth and e-health accessibility.

While existing studies often evaluate physical and digital healthcare networks separately, providing valuable insights into the vulnerabilities and limitations of public services, they do not capture a comprehensive picture of healthcare accessibility. To address this gap, an integrated metric that simultaneously considers

physical access to healthcare facilities and digital access to e-health services is proposed in the present paper.

The rest of the document is organized as follows, Sect. 2 introduces the proposed methodology for physical and virtual accessibility. Section 3 describes the results of the applied methodology on South Africa, while Sect. 4 presents the discussion of the results. Finally Sect. 5 concludes our work and suggests new research avenues.

2 Methodology

In the present section, the methodology to measure physical and digital accessibility is introduced. The first step is the *Data gathering* phase. In this phase we combine the measurements of physical [13] and digital [11] accessibility into a H3 resolution level 8 (*i.e.*, $H3_8$) grid containing: (1) the ID of the $H3_8$ where the measurements were taken (ID_O), (2) the ID of the $H3_8$ where the preferred healthcare facility is located (ID_D), (3) the physical accessibility measurement, which is the time to go from of the ID_O to ID_D, (4) upload speed of the ID_O in Mbps, (5) download speed of the ID_O in Mbps, and (6) latency of the ID_O as shown in Table 1.

Table 1. Data Example

ID_O	ID_D	Time	Upload	Download	Latency
8928308280fffff	8928308283fffff	17	75	50	0.95
8928312380fffff	8928312383fffff	27	35	70	0.15
...	...	...	...	...	...
8928308155fffff	8928308945fffff	31	90	70	0.05

Once the data is collected, a *Data Imputation* process is performed to complete the missing data in the hexagonal grid. For this task, we rely on Inverse Distance Weighting (IDW) [2], which estimates as weighted averages of nearby points based on squared haversine distance, using k-nearest neighbors.

For the *Accessibility measure*, graph-based, composite index, and unsupervised learning models are implemented. A graph model would allow us to represent complex access networks with multiple modes (physical and digital) in a single structure. The first graph approach is a multiplex network including both physical and digital connections between nodes representing H3 hexagons. Each connection is weighted with a physical and digital cost. The former costs are based on normalized Dijkstra travel times, and the latter costs are computed using the inverse of download and upload speeds plus latency. Both cost types are normalized to be between 0 and 1 using the 99th-quantile to reduce the impact of outliers. Digital edges are only included if minimum performance thresholds are met (e.g., bandwidth and latency). Costs are then scaled by parameters α

and β. From these two connections, we build a directed multigraph. Using Dijkstra's algorithm, we compute shortest paths based on combined costs, always selecting the lower-cost edge. This allows us to analyze mode choices and how access differs with different β values (*c.f.*, Algorithm 1).

Algorithm 1. Construct Multi-Edge Graph

Require: Dataset $\mathcal{D}$ with (i, j) pairs, parameters α, β, thresholds $D_{\min}, U_{\min}, L_{\max}$
Ensure: Empty Directed Multigraph G
for all $(i, j) \in \mathcal{D}$ **do**
  Obtain c_{ij}^{phys} and $c_{ij}^{\text{dig}\beta}$
  Add edge $(i \to j)$ with weight $c_{ij}^{\text{phys}\alpha}$ and mode = `physical`
  if $d_{ij} \geq D_{\min}$ and $u_{ij} \geq U_{\min}$ and $l_{ij} \leq L_{\max}$ **then**
    Add edge $(i \to j)$ with weight c_{ij}^{dig} and mode = `digital`

In this variant, we use a standard directed graph with separate physical and digital nodes to separate physical and digital access. Each node has a digital counterpart, and mode switching is modeled explicitly via "switch" edges with an optional activation cost. As before, the costs are normalized using quantiles to reduce outliers. Physical cost is based on travel time, digital cost on the inverse of bandwidth and on latency. Digital edges are added only if quality thresholds are met. This design avoids multiple edges between nodes under consideration while allowing fine-grained control over mode switching. Although it nearly doubles the graph size, it offers greater transparency and flexibility (*c.f.*, Algorithm 2).

Algorithm 2. Build Multi-Layer Graph with Mode-Switching

Require: Normalized DataFrame D, thresholds for bandwidth and latency
Require: Switching costs $c_{\text{switch}}^{\text{origin}}, c_{\text{switch}}^{\text{dest}}$
Initialize directed graph G
for all row $r \in D$ **do**
  Define nodes o, d and digital counterparts o_g, d_g
  Compute C_{ij}^{phys} and C_{ij}^{dig}
  Add edge $(o \to d)$ with weight C_{ij}^{phys} and $(o_g \to d_g)$ with weight C_{ij}^{dig}
  Add edge $(o \to o_g)$ with weight $c_{\text{switch}}^{\text{origin}}$ and $(d_g \to d)$ with weight $c_{\text{switch}}^{\text{dest}}$
  if $t_{total} > \tau$ and latency thresholds are met **then**
    Add edge $(o_g \to d_g)$ with weight C_{ij}^{dig}

Regarding the composite index, we rely on the assignment of weights to reflect the relative importance of each variable ($\omega_1 + \omega_2 = 1$) [8]. Thus, the metric combines physical and digital accessibility using weights into a normalized score, using inverse travel time scaled by the 99th percentile. Finally, the clustering process groups regions with similar accessibility profiles without relying on arbitrary thresholds, employing two distinct approaches. Approach 1 uses K-Means

clustering on the digital accessibility score and scaled travel time in the H3 cells as input. Three clusters ("Sufficient," "Limited," and "Insufficient") were determined based on the elbow method and silhouette scores. In contrast, Approach 2 expands the feature set to include travel time, upload speed, download speed, and latency, applying K-Means with five clusters ("Sufficient", "Moderate", "Limited", "Poor", and "Insufficient"). Like Approach 1, cluster counts and labels were informed by validation metrics. The key differences lie in the number of features (2 vs. 4), cluster granularity (3 vs. 5), and the explicit incorporation of raw digital metrics (speeds/latency) in Approach 2, offering a more nuanced segmentation of healthcare accessibility. The last stage of the process is the *map visualization* to build cartographies summarizing the results to better understand combined physical and digital accessibility.

3 Results

In this section, we present the result of applying the proposed methodology over the physical and digital accessibility data from South Africa, which offers a valuable case for studying healthcare access in middle-income countries [14]. While urban areas are fairly well-covered with healthcare facilities, rural regions often lack adequate services. With about 79% of the population having some form of internet access [7], the country presents an opportunity to explore integrated physical and digital healthcare accessibility.

For the present effort, two main datasets about physical and digital accessibility, in South Africa, were gathered from prior work by the World Bank and Stanford Xplore. The former dataset includes approximately 885,000 modeled travel routes between residential locations and healthcare facilities, containing information on travel distance, time, paths, and relevant demographics such as population density and rural/urban classification. The latter dataset contains 1.5 million of quality of service (QoS) measurements such as download speed, upload speed, and latency at national level. Additionally, supplementary datasets were incorporated to support the network analysis component. These include a detailed road network of approximately 4 million edges, corresponding road nodes approximately 2.9 million nodes, and a national population dataset.

Our exploratory analysis identified limitations in the QoS metrics for landline and mobile datasets. To facilitate the integration of physical and digital accessibility metrics, we aligned each journey origin from the healthcare travel dataset with its corresponding H3 cell at the same resolution, resulting in a dataset comprising over 800,000 journeys from more than 170,000 unique origin points—all annotated with population data and potential internet speed and latency values. To reduce missing values of QoS mobile and landline information were merged using min/max operations on speed and latency. Given these data constraints, our analysis is best understood as a conceptual framework for integrating physical and digital healthcare access, illustrating the potential of such an approach. To estimate missing data of download speed, upload speed, and latency, we tried different method from geostatics (Ordinary Kriging), supervised techniques

(XGBoost) and unsupervised methods (K-Means clustering, IDW, and Kernel Density Estimation). To evaluate the best imputation method, Jensen-Shannon Divergence (JSD), was used. Thus, JSD measures the similarity between probability distributions in a symmetric and bounded way. Based on the results in Table 2, IDW at resolution eight follows the best the data distribution.

Table 2. Summary of Jensen-Shannon Divergence (JSD) mean values for different imputation methods (KDE, IDW, XGBoost, Kriging, and K-Means) evaluated across various holdout sets (H3) with and without urban/rural data splitting. Lower JSD values indicate better similarity between imputed and original data distributions.

Method	H3	Split	JSD mean	Method	H3	Split	JSD mean
KDE	7	✓	0.026	XGBoost	8	✓	0.316
KDE	6	✓	0.037	XGBoost	8	✗	0.333
KDE	7	✗	0.042	Kriging	7	✗	0.28
KDE	6	✗	0.055	Kriging	6	✓	0.323
IDW	8	✓	0.065	Kriging	7	✓	0.325
IDW	8	✗	0.006	Kriging	6	✗	0.329
IDW	7	✓	0.073	Kriging	8	✓	0.379
IDW	7	✗	0.074	Kriging	8	✗	0.405
IDW	6	✗	0.084	K-Means	8	✗	0.383
IDW	6	✓	0.091	K-Means	7	✗	0.397
XGBoost	6	✓	0.113	K-Means	7	✓	0.397
XGBoost	6	✗	0.183	K-Means	6	✗	0.418
XGBoost	7	✓	0.234	K-Means	6	✓	0.422
XGBoost	7	✗	0.263	K-Means	8	✓	0.441

Table 2 presents a summary of the Jensen-Shannon Divergence (JSD) averages obtained for different imputation methods, including Kernel Density Estimation (KDE), Inverse Distance Weighting (IDW), XGBoost, Kriging, and K-Means. The table reports the JSD mean values across different holdout sets (denoted as H3 with values 6, 7, and 8) and under conditions categorized as either urban/rural data split (✓) or no split (✗).

Results indicate that KDE and IDW methods generally achieve lower JSD mean values compared to the other techniques, suggesting a better performance in approximating the underlying data distributions, especially when the data are split. Notably, KDE with split and H3 values 7 and 6 yielded the lowest JSD means of 0.026 and 0.037, respectively. Conversely, K-Means demonstrated the highest JSD means across most configurations, with values reaching up to 0.441, indicating comparatively poorer imputation quality. Kriging and XGBoost exhibited intermediate performance, with Kriging notably achieving lower JSD when the data was split and H3 values were 6 and 7, whereas XGBoost showed higher JSD means overall. These findings illustrate the varying efficacy of the

imputation methods in minimizing divergence between the imputed and true data distributions, highlighting KDE and IDW as more effective approaches under the tested con

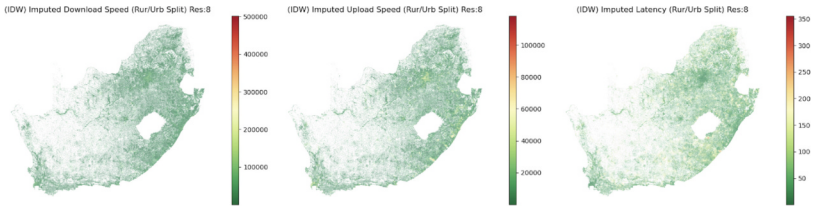


Fig. 1. Imputed download speed, upload speed, and latency across South Africa, based on a hexagonal grids of resolution 8. Imputation is performed separately for urban and rural hexagons.

Figure 1 presents spatial interpolations of imputed download speed, upload speed, and latency across South Africa, each mapped according to a hexagonal grid of resolution 8. The figure displays results generated using an imputation strategy that separates the data into urban and rural categories before performing interpolation, thus addressing systematic differences in network connectivity between these regions. Each panel in the figure visualizes a different metric: the first panel shows the distribution of imputed download speeds, the second presents upload speeds, and the third illustrates latency. The color gradients in each map represent the variation in the respective metric, with higher values shown in darker shades and lower values in lighter shades. This partitioned imputation approach allows for more accurate representation of local connectivity by accounting for the heterogeneity in service quality between urban and rural areas.

Once the dataset is in the H3 grid and complete, the physical and digital accessibility metric is computed. The first method rely on a *Multiplex Network* including both physical and digital connections between nodes. Figure 2 depicts the distribution of path types by β value, based on shortest-path computations in the Multi-Edge Graph model. In this graph, each pair of nodes is connected by both a physical and a digital edge, with associated costs reflecting travel time and network performance, respectively. The parameter β scales the influence of digital cost relative to physical cost in the total path calculation. As β increases, digital access becomes more costly, leading to a shift toward paths dominated by physical links. This figure illustrates how the balance between digital and physical modalities changes as a function of β, showing an increasing dominance of physical paths under higher parameter weights. The parameter β modulates the relative influence of digital cost versus physical cost in determining the optimal path. As β increases from 0.2 to 1.0, representing an increasing penalty on digital access costs, there is a clear shift in the path distribution from predominantly digital to predominantly physical. Specifically, at lower β values, digital paths

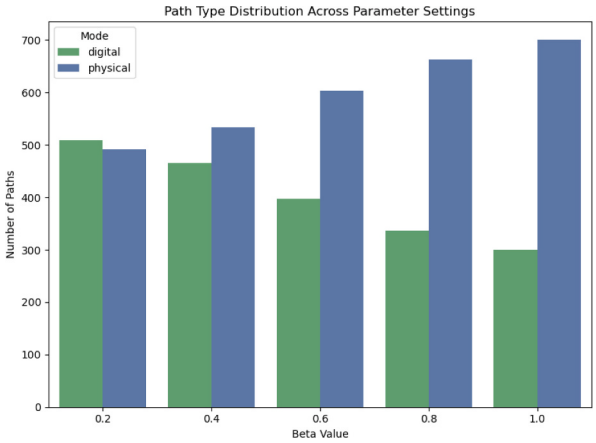


Fig. 2. Distribution of digital and physical path counts as a function of the parameter β in the Multiplex Graph model. Increasing β corresponds to greater penalty on digital access costs, resulting in a progressive shift from digital-dominated shortest paths to those primarily composed of physical links.

slightly outnumber physical ones; however, as β rises, the total number of physical paths increases significantly, while digital paths correspondingly decrease. This trend illustrates how increasing the cost sensitivity to digital links redirects optimal routing toward physical networks, emphasizing the model's ability to capture dynamic modality preference dependent on cost weighting.

For the second method a *Multi-Layer Graph Model* with a standard directed graph framework is employed to distinguish between physical and digital access, utilizing distinct nodes for each modality. Every physical node is paired with a corresponding digital node, and the transition—or "mode switching"—between these modalities is explicitly represented through dedicated "switch" edges, which can include an optional activation cost to capture transition dynamics in the network. This modeling approach enables a nuanced analysis of multimodal access pathways, reflecting both direct and switched routes in the context of integrated service provision.

Figure 3 presents violin plots depicting the distributions of path costs in a Multi-Layer Graph model, separated by mode (physical vs. digital) and levels of switching cost (0.0 and 0.1). On the left, at a switch cost of 0.0, both physical and digital modes exhibit relatively lower median path costs, with digital paths skewing slightly lower, indicating efficient access via either modality when switching penalties are absent. On the right, when the switch cost increases to 0.1, the cost distribution for digital paths shifts upward, demonstrating higher median and overall path costs. This shift underscores a reduced preference for digital routes as switching costs rise, resulting in a relatively greater reliance on physical modes. The distribution for physical paths remains more stable, highlighting their robustness to increased switching penalties. As switching costs increase,

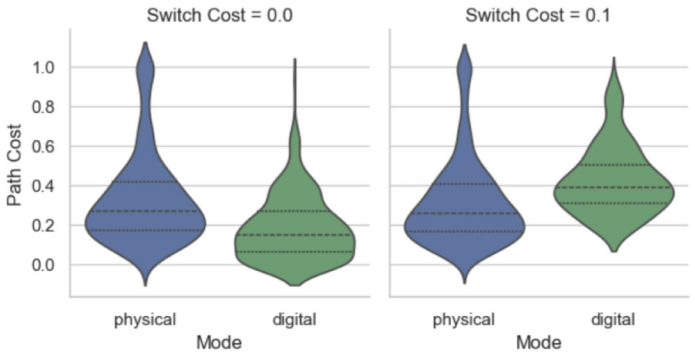


Fig. 3. Violin plots of path cost distributions in a Multi-Layer Graph, stratified by mode (physical vs. digital) and switch cost. Increased switching costs elevate digital path costs and increase reliance on physical routes, illustrating the impact of mode-switching penalties on multi-modal network access patterns.

the distribution of digital path costs shifts upward, indicating reduced attractiveness of digital routes and a relative increase in reliance on physical paths. This illustrates how mode-switching penalties influence access patterns in this mixed-modality network. These results quantitatively illustrate how even modest mode-switching penalties can reshape accessibility patterns in multi-modality networks by discouraging transitions to digital pathways.

The next approach is to build a composite index [8] with physical and digital metrics. Thus, Fig. 4 presents the spatial distribution of origin nodes based on a composite total access score for South Africa. This score is calculated from a weighted combination of two key parameters, namely physical access (weight = 0.6) and digital access (weight = 0.4). The map uses a color gradient ranging from red (low access) to green (high access) to visually represent the relative accessibility of each region.

The map reveals substantial spatial heterogeneity in access levels across the country. Green-shaded areas indicate regions with relatively high overall access, suggesting better infrastructure and connectivity. In contrast, regions shaded in red highlight areas with limited access, potentially reflecting infrastructural deficits or aspects of the digital divide. The greater weighting of physical access (60%) emphasizes the crucial role of tangible infrastructure in shaping overall accessibility.

The spatial distributions depicted in Figs. 4 and 5 provide complementary insights into the geographic patterns of accessibility and population density across South Africa. Figure 4 illustrates the origin nodes based on a composite total access score, where the gradient from 0.1 to 0.8 reflects varying levels of accessibility, with higher scores likely indicating better-connected or more central locations. In contrast, Fig. 5 employs a log-transformed population density map, using a color gradient from red (low density) to green (high density) to highlight stark regional disparities, with urban centers exhibiting pronounced

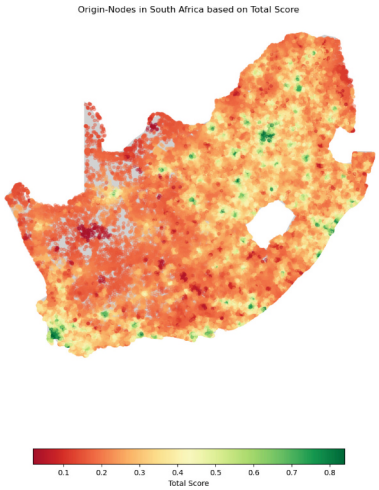


Fig. 4. Spatial distribution of origin nodes in South Africa based on a composite total access score (range: 0.1–0.8). Higher scores indicate greater accessibility, revealing regional disparities in connectivity and infrastructure. (Color figure online)

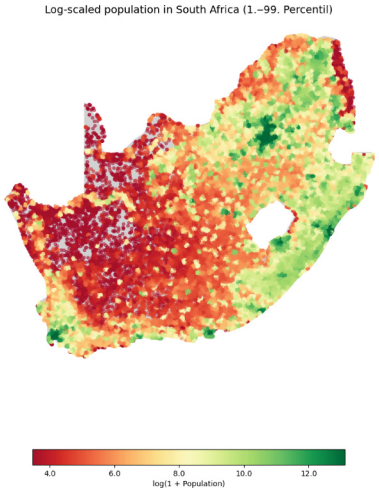
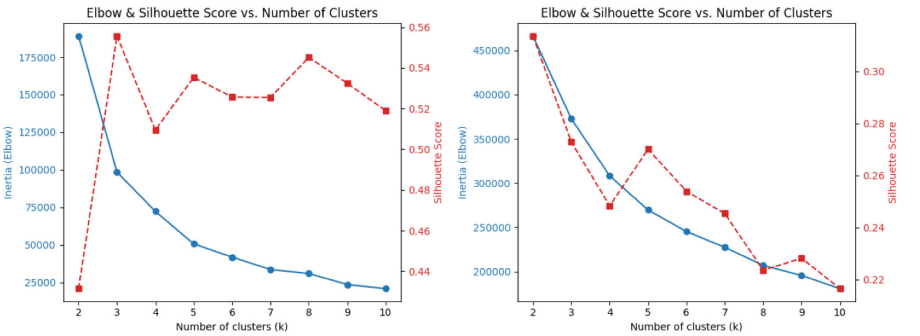


Fig. 5. Log-scaled population density distribution across South Africa (1stâĂŞ99th percentile). The gradient (red to green) highlights urban-rural divides, with dense urban clusters (green) contrasting with sparsely populated areas (red). (Color figure online)

density compared to rural areas. A comparative analysis suggests potential correlations between high-accessibility nodes (Fig. 4) and densely populated regions (Fig. 5), as urban centers often coincide with robust infrastructure and connectivity. Together, these visualizations underscore the spatial heterogeneity in both human settlement and infrastructure development within South Africa.

A key challenge for the aforementioned technics lies in assigning weights to digital and physical access modes, which inherently imposes assumptions about what constitutes better access. Without input from affected communities, this risks misrepresenting local needs. To tackle this challenge, an unsupervised K-means clustering approach was tested, allowing the data to reveal patterns without relying on externally defined preferences. The clusters were tagged as "Sufficient", "Moderate", "Limited", "Poor", and "Insufficient" to indicate the level of healthcare accessibility.



(a) Approach 1 (b) Approach 2

Fig. 6. Elbow method (inertia) and silhouette score as a function of the number of clusters (k) for two clustering approaches. Approach 1 (a) shows higher inertia values but a clear peak in silhouette score at lower k, while Approach 2 (b) exhibits lower inertia and more stable silhouette performance, suggesting different optimal cluster counts for each method.

To find the best K for both clustering approaches, the elbow and silhouette scores were computed for different approaches as shown in Fig. 6. The figures illustrate the relationship between the number of clusters (k) and two key metrics—inertia (elbow method) and silhouette score—for two distinct clustering approaches. Figure 6a depicts the **Approach 1**, where inertia decreases sharply from 450,000 to 200,000 as the number of clusters increases from 2 to 10, suggesting a gradual improvement in cluster compactness. Meanwhile, the silhouette score peaks at 0.56 for a lower number of clusters (k=2) and declines to 0.22 as k increases, indicating that fewer clusters yield better-defined groupings. Figure 6b shows the *Approach 2* exhibiting a similar trend in inertia, dropping from 175,000 to 25,000 across the same range of k values, but the silhouette score remains relatively stable around 0.50 for smaller k before declining, reinforcing the preference for a modest number of clusters. Collectively, the results highlight a trade-off between minimizing inertia and maximizing silhouette score, with optimal performance typically achieved at intermediate k values.

To assess the accessibility of healthcare facilities, we calculate the percentage of South Africans who reside in regions belonging to each cluster. The results

are shown in Table 3. *Approach 1* reveals that about 51.1% of the South African population has "Sufficient" access, whereas 47.1% of the population has "Limited" access to healthcare. Approach 2 provides a more granular categorization of healthcare accessibility. As per the results of *Approach 2*, 27.6%, 32.4%, 26.9%, and 12.6% of the South African population has "Sufficient", "Moderate", "Limited", and "Poor" access respectively. Across both approaches, we find that only a small portion of the population are categorized as having "Insufficient" access.

Table 3. Percentage of the South African population with their respective categorized level of healthcare accessibility.

Category	Approach 1	Approach 2
Sufficient	51.1%	27.6%
Moderate	–	32.4%
Limited	47.1%	26.9%
Poor	–	12.6%
Insufficient	1.8%	0.5%

Figures 7 and 8 present a visual representation of the clustering across the regions of South Africa for Approaches 1 and 2 are respectively. The maps use colors ranging from red (low access) to green (high access) to visually represent the relative accessibility of each region. Similar to the results in Fig. 4, the maps reveal substantial spatial heterogeneity in access levels across the country.

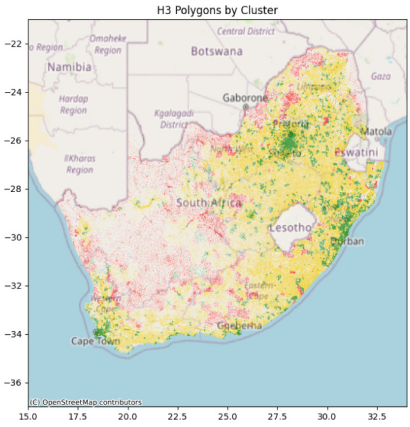


Fig. 7. Spatial classification of healthcare access in South Africa using Approach 1. H3 hexagons are categorized into three levels—'Sufficient' (green), 'Limited' (yellow), and 'Insufficient' (red)—reflecting regional disparities in healthcare availability. (Color figure online)

Figure 7, generated using *Approach 1*, categorizes regions into three distinct classes: "Sufficient," "Limited," and "Insufficient" access to healthcare, represented by green, yellow, and red polygons, respectively. This simplified classification provides a broad overview of healthcare accessibility, highlighting areas with critical deficiencies in red, while regions with adequate services are marked in green. The intermediate yellow regions suggest zones where access is neither fully sufficient nor entirely lacking, indicating potential targets for targeted improvements.

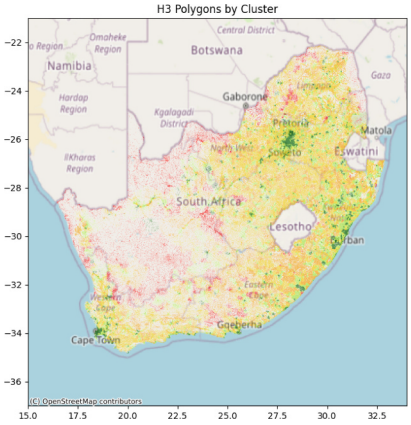


Fig. 8. Refined spatial classification of healthcare access in South Africa using Approach 2. H3 hexagons are segmented into five tiers—'Sufficient' (dark green), 'Moderate' (light green), 'Limited' (yellow), 'Poor' (orange), and 'Insufficient' (red)—providing a granular assessment of accessibility challenges. (Color figure online)

In contrast, Fig. 8, produced using *Approach 2*, offers a more granular classification system with five categories: "Sufficient," "Moderate," "Limited," "Poor," and "Insufficient," depicted by dark green, light green, yellow, orange, and red polygons, respectively. This refined stratification allows for a more nuanced understanding of healthcare access disparities. The inclusion of intermediate categories such as "Moderate" and "Poor" enables finer distinctions between regions, revealing gradients of accessibility that may inform prioritized interventions. The spatial patterns in both figures likely reflect underlying socioeconomic, infrastructural, or geographic determinants, with the darker hues indicating more severe challenges in healthcare provision. Together, these visualizations underscore the variability in healthcare access across South Africa and emphasize the importance of methodological choices in spatial analysis for policy planning.

4 Discussion

In the preset section, we discuss about the advantages and disadvantages of the tested methods to measure physical and digital accessibility as a single accessi-

bility metric. Regarding the *multi-edge graph* model, it offers a robust framework for representing physical and digital healthcare access simultaneously by incorporating multiple weighted connections between nodes based on travel time and network QoS. This model allows dynamic path selection through tunable parameters, enabling scenario analyses that reflect various infrastructure priorities. However, its outcomes are sensitive to parameter settings, which require careful empirical calibration, and the model's complexity due to the dual connections for each node increases computational demands and may limit scalability. The *multi-layer graph* model separates physical and digital modes by pairing each physical node with a corresponding digital node and introducing switch edges to model transitions between access types. This explicit structure enhances control and transparency, permitting detailed sensitivity analyses. Although this model improves interpretability by reducing redundant edges. As downside, it demands precise data normalization and empirically grounded parameters, and may oversimplify the complexity of hybrid or sequential access paths.

The composite index approach simplifies accessibility assessment by aggregating normalized physical and digital metrics into a weighted score, which facilitates easy spatial comparisons and can be tailored to reflect policy priorities through adjustable weighting. Despite its straightforwardness and intuitive communication value, this method relies on subjective weight assignments that might not align with actual user experiences and risks obscuring spatial or demographic nuances by aggregating diverse factors.

Unsupervised K-means clustering provides a data-driven way to categorize areas based on similar accessibility profiles without imposing pre-set thresholds, enabling classification for targeted interventions. This approach avoids policy bias, letting patterns naturally emerge from the data. However, cluster interpretability can be challenging for non-experts, the determination of optimal cluster numbers requires empirical validation, and cluster assignments can be unstable with changes in input data, complicating ongoing tracking and policy application.

In summary, each method presents a distinct balance of complexity, interpretability, and data requirements: multiplex and multi-layer graphs offer detailed network modeling but with higher computational and data demands; composite indices facilitate simplicity but with limitations in capturing multimodal dynamics; and clustering provides flexible, objective categorization while facing challenges in cluster interpretation and stability. Given the task of integrating both physical and digital healthcare accessibility for policy, and planning the clustering-based approach emerges as the most suitable method.

5 Conclusion

In the present effort, we have compared IDW, KDE, XGBoost, Kriging, and K-Means imputation methods. Additionally, we have tested graph based, weight assignment, and unsupervised based methods to combine physical and digital accessibility to represent in a single accessibility metric. We showed that IDW

imputation technic together with cluster based methods are more suitable and understandable for this task. As new research avenues, we would like to take into account the time dimension to measure accessibility over time taking into account traffic jams and network congestion.

References

1. Anawade, P.A., Sharma, D., Gahane, S., Anawade Sr, P.A., Sharma, D.S.: A comprehensive review on exploring the impact of telemedicine on healthcare accessibility. Cureus **16**(3) (2024)
2. Benmoshe, N.: A simple solution for the inverse distance weighting interpolation (IDW) clustering problem. Sci **7**(1), 30 (2025)
3. Chowdhury, J., Ravi, R.P.: Healthcare accessibility in developing countries: a global healthcare challenge. J. Clin. Biomed. Res. **152**(4), 2–5 (2022)
4. Coombs, N.C., Campbell, D.G., Caringi, J.: A qualitative study of rural healthcare providers' views of social, cultural, and programmatic barriers to healthcare access. BMC Health Serv. Res. **22**(1), 438 (2022)
5. Geracitano, J., Barron, L., McSwain, D., Khairat, S.: How is digital health suitability measured for communities? A systematic review. Digital Health **10**, 20552076241288316 (2024)
6. Levesque, J.F., Harris, M.F., Russell, G.: Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int. J. Equity Health **12**(1), 18 (2013)
7. Masia, L.: South Africa connect (2025). https://www.gov.za/blog/south-africa-connect. Accessed 22 July 2025
8. Mazziotta, M., Pareto, A., et al.: Methods for constructing composite indices: one for all or all for one? Rivista italiana di economia, demografia e statistica **67**, 67–80 (2013)
9. Mseke, E., Jessup, B., Barnett, T.: Impact of distance and/or travel time on healthcare service access in rural and remote areas: a scoping review. J. Transport Health **37**, 101819 (2024)
10. Omaghomi, T.T., Arowoogun, J.O., Akomolafe, O., Odilibe, I.P., Elufioye, O.A.: Telemedicine in rural Africa: a review of accessibility and impact. World J. Adv. Res. Rev. **21**(2), 421–431 (2024)
11. Nunez-del Prado, M., Barrera, J., Klaiber, C., Tariverdi, M.: Mapping digital accessibility: a machine learning framework for improving e-health service delivery (2025)
12. Putra, H.N., Zani, B.N., Hasyim, D.M., Muntasir, M., et al.: e-health and digital transformation in increasing accessibility of health services. J. World Future Med. Health Nurs. **2**(1), 119–132 (2024)
13. Tariverdi, M., Nunez-Del-Prado, M., Leonova, N., Rentschler, J.: Measuring accessibility to public services and infrastructure criticality for disasters risk management. Sci. Rep. **13**(1), 1569 (2023)
14. The World Bank: World bank open data: South Africa and cross territory (za–xt) (2025). https://data.worldbank.org/?locations=ZA--XT. Accessed 22 July 2025

Intelligent Query System for Consumer Protection Regulations in Perú Using Hierarchical Vector Databases

Franco Cárdenas-Martínez and Antonio Arroyo-Paz(✉)

Universidad Nacional de San Agustín, Arequipa, Peru
{fcardenasm,aarroyop}@unsa.edu.pe

Abstract. This paper presents the design and implementation of an intelligent query system for the Consumer Protection and Defense Code of Peru (Law No. 29571). The system enables users to ask questions in natural language and receive legally grounded answers. A hierarchical vector database is employed, structured into chapters, subchapters, articles, and content, and indexed using semantic embeddings. The implementation integrates technologies such as ChromaDB with a SQLite backend and a Retrieval-Augmented Generation (RAG) pipeline, leveraging the GPT-4o-mini model, which was selected for its high efficiency and low token consumption.

During validation, 60 representative queries were evaluated using the LLM-as-a-Judge technique, with faithfulness and answer relevancy metrics assessed via DeepEval. A comparative analysis among generative models (GPT-4o-mini, Claude 3 Haiku, and GPT-3.5-turbo) revealed that GPT-4o-mini provided the best balance between computational cost and semantic performance . The results confirm the system's effectiveness in making legal information accessible to users without legal expertise. This approach contributes to the democratization of legal knowledge in everyday scenarios.

Keywords: Legal query system · Vector database · Natural language processing · GPT-4o-mini · ChromaDB · Consumer protection · LLM-as-a-Judge · RAG · Token efficiency

1 Introduction

In Peru, consumer rights are regulated by Law No. 29571, known as the *Consumer Protection and Defense Code*, and overseen by the National Institute for the Defense of Competition and Protection of Intellectual Property (INDECOPI) [1]. This legal framework comprises numerous chapters, articles, and complementary provisions, making it highly complex. Moreover, the dispersion of consumer-related norms across general laws, sectoral regulations, and administrative jurisprudence further complicates its understanding and accessibility, especially for citizens, micro-entrepreneurs, and users without legal training.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 125–136, 2026.
https://doi.org/10.1007/978-3-032-20322-9_9

Traditional legal search systems—primarily based on keyword matching or exact patterns—are limited in handling queries that require semantic or contextual interpretation [2,3]. Additionally, the technical language used in legal texts often acts as an exclusionary barrier for non-specialists, restricting their ability to assert the rights guaranteed by the normative framework. In fact, only 54.6% of consumers in Peru have heard about the existence of protective legislation such as the Consumer Code [4].

To address these challenges, this paper proposes an intelligent legal query system based on the *Retrieval-Augmented Generation* (RAG) architecture. This system combines Natural Language Processing (NLP) techniques with a hierarchical vector database constructed from the official Consumer Code. The corpus is structured into four hierarchical levels: chapter, subchapter, article, and content. It is further enriched with thematic categorization around key issues such as warranties, contracts, e-commerce, sanctions, and legal deadlines [5].

RAG has proven effective in complex domains where semantic search and generation deliver accurate and context-aware answers [6,7]. In our system, legal fragments are segmented through overlapping chunking techniques and vectorized using semantic embeddings, allowing natural language queries. This optimized segmentation minimizes token usage, improving efficiency in resource-constrained environments.

In terms of operational efficiency, our system using GPT-4o-mini outperforms other models in token consumption. On average, GPT-4o-mini consumes 835.8 tokens per query, compared to 1042.3 tokens with GPT-3.5-turbo and 987.6 tokens with Claude 3 Haiku. This represents a 20–30% advantage in efficiency, making it particularly crucial for applications that involve multiple interactions per session.

The system was evaluated using 60 representative questions covering trivial, hierarchical, and inferential aspects of consumer rights. Reference answers were constructed from the normative text. We employed the DeepEval framework under the *LLM-as-a-Judge* paradigm to assess answer quality, using metrics of faithfulness (adherence to retrieved content) and answer relevancy (alignment with user intent) [8]. Each metric was computed by averaging scores, standard deviations, and proportions exceeding thresholds of 0.80 and 0.85, using GPT-4 (temperature 0.2).

Additionally, our system achieves a notable reduction in token consumption during response generation, improving cost-efficiency and responsiveness—an essential feature in multi-query legal scenarios.

Results show that the system generates informative, grounded, and understandable answers—even for unstructured questions. This solution represents a significant step toward democratizing legal knowledge by enhancing public access to critical normative information.

The objective of this paper is to develop and evaluate a Retrieval-Augmented Generation (RAG)-based system designed to answer natural language queries grounded in the Peruvian Consumer Protection and Defense Code (Law No. 29571). The remainder of the paper is struc-

tured as follows: Sect. 2 reviews related work; Sect. 3 presents the methodology; Sect. 4 outlines the system architecture; Sect. 5 describes the implementation; Sect. 6 discusses the experimental results; and Sect. 7 provides conclusions and future directions.

2 Related Works

Lima et al. [9] proposed Poly-Vector Retrieval, an architecture for legal search that distinguishes between content and reference embeddings, achieving more accurate retrieval in complex regulatory documents such as the Brazilian Constitution. This semantic separation improves the relevance of results without penalizing performance, highlighting its applicability in hierarchical legal environments where referential structure is critical.

Barron et al. [10] integrated vector databases, knowledge graphs, and hierarchical NMF decomposition to build a contextualized legal retrieval system. Their approach enables queries at differentiated levels (statutes, decrees, rulings), and their evaluation shows significant reductions in interpretative errors compared to traditional legal search models, especially in multi-level normative systems.

Sugathadasa et al. [11] presented a legal document retrieval system that combines word and document embeddings trained on specialized legal corpora. Their thorough evaluation on legislative texts demonstrates that vector-based models consistently outperform TF-IDF and BM25 in semantic precision, reinforcing the relevance of dense representations in legal applications where terminological accuracy is essential.

Chalkidis et al. [12] developed Legal-BERT, a pre-trained language model specifically trained on multilingual legal corpora. Their comparative study shows that this domain-specific adaptation substantially improves performance on tasks such as normative classification, clause extraction, and fragment retrieval, outperforming general-purpose models both in semantic accuracy and in reducing hallucinations in complex legal contexts.

Peng and Chen [13] proposed Athena, a specialized system for judicial outcome prediction that integrates a hierarchical vector database with semantic embeddings and a Retrieval-Augmented Generation (RAG) mechanism. Their evaluation on the CAIL2018 dataset showed a 34% accuracy improvement (91% using GPT-4o) over baseline methods, highlighting its effectiveness in reducing hallucinations and managing optimal context windows ($k = 32$).

3 Methods

This section describes the procedures and techniques used to develop the intelligent query system for the Consumer Protection and Defense Code (Law No. 29571).

3.1 Corpus Preparation

The official PDF of the Code was obtained from the INDECOPI portal [1]and stored in the `docs/` directory. Using LangChain's `PyPDFLoader`, content was extracted page by page. A subsequent cleaning and normalization process was applied, which included: removing duplicate page numbers, discarding irrelevant special characters, and standardizing spaces and hyphens in line with recommended practices for legal retrieval systems [13].

3.2 Hierarchical Segmentation and Chunking

To preserve the legal structure, the full text was segmented into four hierarchical levels: chapter, subchapter, article, and content, using specific regular expressions (`CAPITULO_RE`, `SUBCAPITULO_RE`, `ARTICULO_RE`) following the approach of Lima et al. [9]. Each hierarchical fragment was further divided into sub-chunks of up to 400 characters with an overlap of 100 characters, using LangChain's `RecursiveCharacterTextSplitter`. This overlap preserves contextual continuity across adjacent chunks, enhancing semantic retrieval based on neural ranking models [14].

3.3 Vectorization and Indexing

Each sub-chunk was converted into a high-dimensional vector using OpenAIEmbeddings, implementing Dense Passage Retrieval schemes [15,16]. The embeddings and their associated metadata (hierarchical level, fragment type, unique identifier) were indexed in `ChromaDB` with an SQLite backend, leveraging its local persistence and k-NN query optimization capabilities [17]. The `Chroma.from_documents(...).persist()` operation ensured the vector store was immediately available for querying.

3.4 Retrieval-Augmented Generation (RAG) Pipeline

The query component was built using LangChain's `RetrievalQA` class and the `gpt-4o-mini` model. During execution, the system: (1) embeds the user query using the same model as for indexing, (2) retrieves the top-10 most relevant subchunks via cosine similarity, (3) constructs an augmented prompt integrating the query and retrieved fragments, and (4) generates the final answer using the LLM. This flow combines dense retrieval precision with advanced generative capabilities, producing contextualized, legally grounded answers and showing significant improvements in legal QA tasks [6].

3.5 Automated Evaluation

A benchmark of 60 representative questions was designed (20 trivial, 20 hierarchical, 20 inferential). Each query was matched with a reference ("golden") answer curated from the legal text. To assess response quality, `DeepEval` was

applied under the LLM-as-a-Judge paradigm [18], using the metrics *Faithfulness* (fidelity to retrieved content) and *Answer Relevancy* (alignment with question intent). The evaluation ran using GPT-4 (temperature 0.2), calculating for each metric the average, standard deviation, and the percentage of cases exceeding the 0.80 and 0.85 thresholds.

3.6 Model Comparison and Token Efficiency

The same benchmark set was executed across three language models: GPT-4o-mini, GPT-3.5-turbo, and Claude 3 Haiku. For each query, total token usage (prompt + completion) was recorded. GPT-4o-mini averaged 835.8 tokens, GPT-3.5-turbo 1042.3, and Claude 3 Haiku 987.6. This evaluation was performed programmatically to quantify resource efficiency during inference(Fig. 1).

Fig. 1. System Development Workflow

4 System Architecture

This section details the architecture of the system, including the construction of the vector store, query flow, user interface, and overall system architecture.

4.1 Vector Store Construction

The vector store construction begins with uploading the legal PDF document, followed by its hierarchical segmentation into normative fragments. Each fragment is transformed into a vector using semantic embedding models. Finally, the vectors are stored in a persistent vector database to enable contextual similarity search (see Fig. 2).

4.2 Query Flow

The query process starts with a user prompt, which is converted into a semantic vector for similarity search in the vector database. The retrieved legal fragments are combined with context and processed by the GPT-4o-mini model to generate a precise and understandable answer. This process is illustrated in Fig. 3.

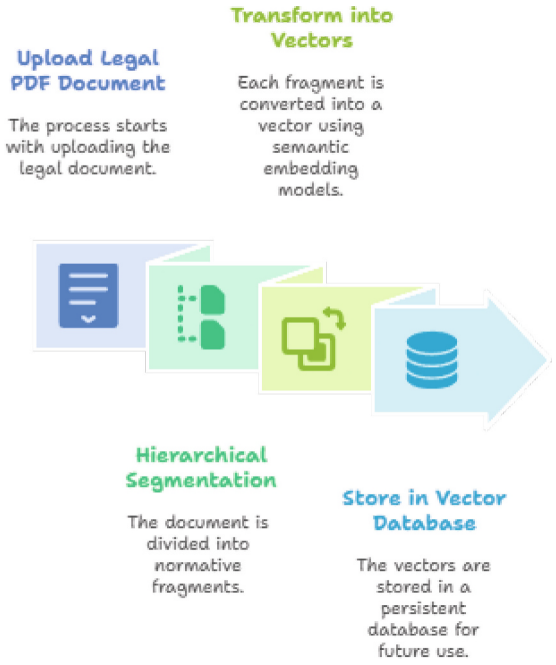


Fig. 2. Flowchart of the Vector Store Construction Process

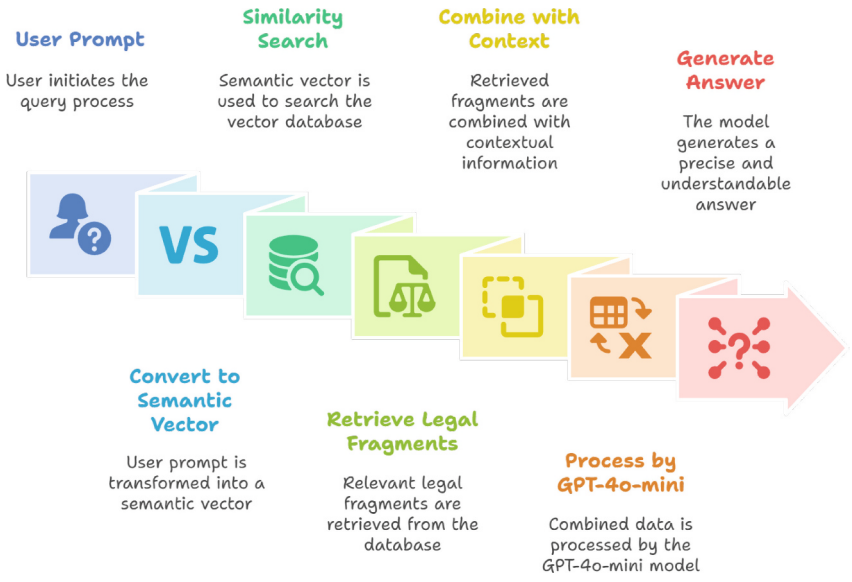


Fig. 3. Flowchart of the Query Process

4.3 User Interface

The system features a web interface built with Streamlit, organized into functional modules: welcome screen, intelligent query, keyword search, document viewer, and FAQ section. This modular organization improves user experience and facilitates navigation across complementary consultation modes (Fig. 4).

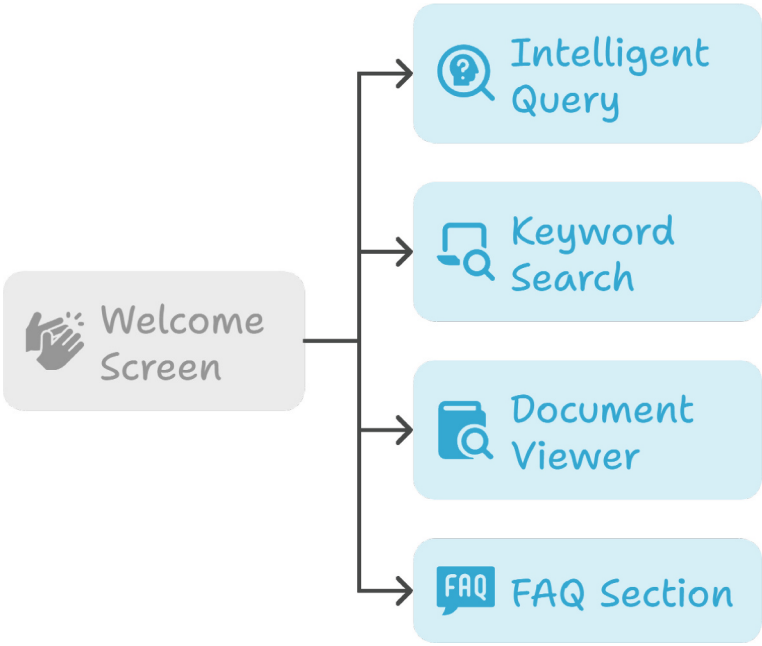


Fig. 4. User Interface Flow

4.4 Overall Architecture

The proposed architecture integrates processing, retrieval, and visualization components in a modular pipeline. During the indexing phase, the system loads the legal document, segments it hierarchically (chapter, subchapter, article), and stores the semantic vectors along with structured metadata in ChromaDB. In the query phase (see Fig. 5), the user submits a natural language question, which is converted to an embedding and compared against the vector store. The most relevant fragments are retrieved and provided as context to GPT-4o-mini, which generates an adapted answer. This output is displayed through a modular Streamlit interface to enhance legal navigation experience.

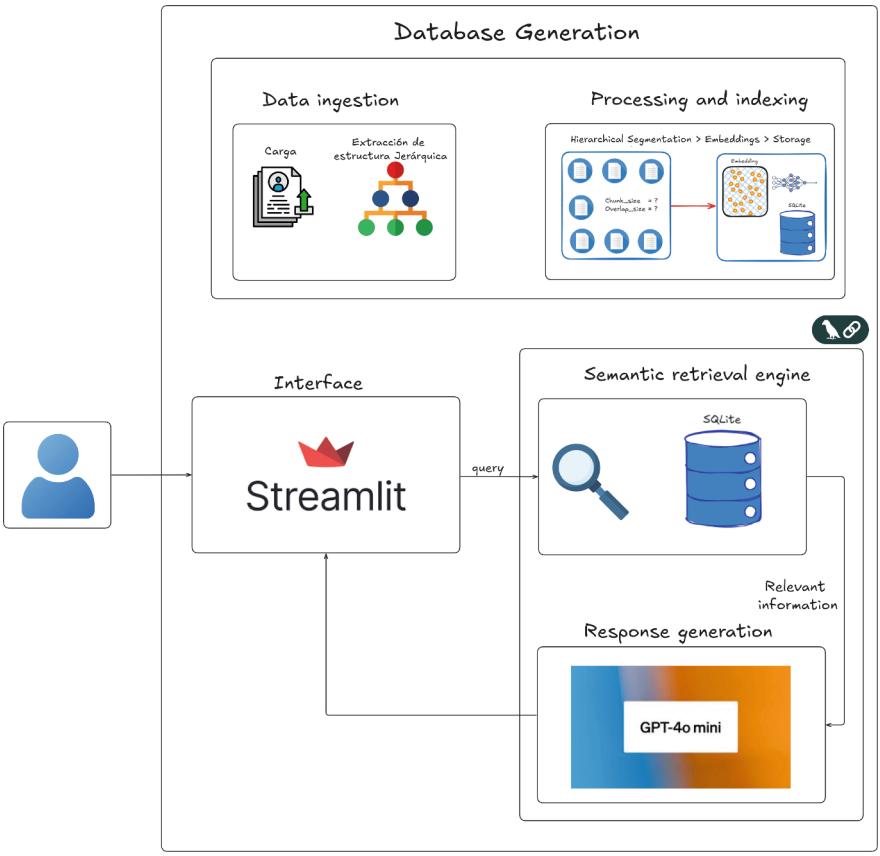


Fig. 5. Overview of the Proposed System Architecture

5 Results and Discussion

To evaluate system performance, a set of 60 test cases was designed and balanced across three categories:

- **Trivial**, to validate operational information of the system (e.g., "What are you?").
- **Hierarchical**, to verify structured extraction (e.g., "Where is misleading advertising regulated?").
- **Inferential**, requiring more complex semantic reasoning (e.g., "What rights do I have as a consumer?").

Each question was paired with a reference ("golden") answer strictly based on the legal text. The RAG pipeline (ChromaDB + OpenAIEmbeddings + GPT-4o-mini) generated responses evaluated using DeepEval under the LLM-as-a-Judge paradigm with GPT-4 (temperature 0.2), based on two key metrics:

- **Faithfulness**, which measures how well the response reflects the retrieved content.
- **Answer Relevancy**, which judges whether the response correctly addresses the intent of the question.

The evaluation relied on two metrics designed for large language model (LLM) output assessment: *Faithfulness* and *Answer Relevancy*. Their mathematical definitions are shown below.

$$\text{Faithfulness}(R, E) = \frac{|\{s \in R \mid s \text{ is supported by } E\}|}{|\{s \in R \mid s \text{ is a factual statement}\}|} \tag{1}$$

$$\text{Answer Relevancy}(R, Q, A) = \frac{|\{k \in K_A \mid k \in R\}|}{|K_A|} \tag{2}$$

where:

- R: Model-generated response.
- E: Retrieved evidence from the vector database.
- Q: Input user query.
- A: Reference answer (gold standard).
- s: A factual statement in R.
- K_A: Set of key elements present in A.

Table 1. Automatic evaluation of 60 cases using DeepEval (GPT-4, temperature 0.2).

Metric	Threshold	Average	Standard Deviation	% ≥ Threshold
Faithfulness	0.81	0.85	0.129	78%
Answer Relevancy	0.85	0.82	0.168	62%

Quantitative results (Table 1) show that the system achieved an average Faithfulness score of 0.85, with 47 out of 60 responses (78%) faithfully reflecting the retrieved content, validating the effectiveness of hierarchical indexing and overlapping chunking. In contrast, the Answer Relevancy metric scored an average of 0.82, with 37 cases (62%) exceeding the 0.85 threshold, suggesting opportunities for improvement in contextual precision. This gap between faithfulness and relevancy indicates that while the system retrieves and cites normative content accurately, responses may sometimes be too generic or not fully aligned with the query focus.

In terms of operational efficiency, GPT-4o-mini demonstrated superior token usage (Table 2), consuming an average of 835.8 tokens per query compared to 1042.3 for GPT-3.5-Turbo and 987.6 for Claude 3 Haiku. This 20âĂŞ30% efficiency advantage is crucial in applications requiring multiple interactions per session.

From a qualitative perspective, three main response patterns were observed:

Table 2. Token consumption comparison per query across models.

Model	Average Tokens	Max Tokens	Min Tokens
GPT-4o-mini	835.8	1062	291
GPT-3.5-Turbo	1042.3	1310	370
Claude 3 Haiku	987.6	1255	345

- **Perfect Responses:** Trivial and well-scoped hierarchical queries (e.g., "How many articles does the Code have?") were answered with exact figures and direct citations.
- **Partial Responses:** Moderately inferential queries (e.g., "What is an abusive clause?") provided correct definitions but lacked precise normative examples.
- **Generic Responses:** Broad queries (e.g., "What rights do I have as a consumer?") tended to produce lists of rights without referencing specific articles.

6 Future Work

To enhance the legal relevance and semantic precision of the system, we identify several areas for future work:

1. **Optimization of vector retrieval:** Adjust the number of nearest neighbors (k) retrieved to capture broader semantic context without sacrificing precision.
2. **Semantic re-ranking:** Implement a re-ranking mechanism based on thematic alignment to prioritize the most contextually relevant fragments.
3. **Prompt engineering improvements:** Refine GPT-4o-mini prompts by incorporating explicit instructions to cite relevant chapters or articles from the legal corpus.

In addition, the current evaluation methodology—relying on a manually crafted reference set and LLM-based scoring—can be complemented in future stages by expert validation from legal professionals. This would ensure greater interpretability and institutional trust. Future iterations may also explore multilanguage support, expansion to other legal domains, and deployment in mobile or offline environments to increase accessibility across diverse user groups.

7 Conclusions

This work presents a smart legal consultation system based on a Retrieval-Augmented Generation (RAG) architecture that integrates hierarchical indexing and semantic embeddings via ChromaDB. The system demonstrates strong effectiveness in representing the complex structure of the Peruvian Consumer Protection Code. Quantitative evaluation revealed high factual accuracy (*Faithfulness*

= 0.85), with 78% of responses accurately reflecting the retrieved legal content. These results validate the hierarchical chunking strategy with context-preserving overlap, enabling precise grounding of responses in the source text—an important advancement for automated legal information systems.

However, the contextual relevance (*Answer Relevancy* = 0.82) highlights certain limitations, particularly in broad or inferential queries, where responses tended to be generic and lacked explicit normative citations. This underscores the need to enhance the retrieval process through semantic re-ranking and prompt engineering techniques that explicitly instruct the model to reference relevant chapters or articles. Such refinements are necessary to improve real-world applicability in citizen-facing legal scenarios.

In terms of operational efficiency, GPT-4o-mini exhibited a significant advantage, with an average token usage of 835 per response—outperforming GPT-3.5-turbo and Claude 3 Haiku by 15–30%. This efficiency in computational cost positions GPT-4o-mini as a practical choice for scalable deployment in legal consultation platforms where resources are limited and multi-turn interactions are frequent.

While the evaluation methodology was effective in quantifying performance, it also carries inherent limitations. The reliance on manually curated reference answers introduces the potential for interpretative bias, and the use of LLMs (GPT-4) as evaluators may inherit systemic biases from the underlying models. Therefore, integrating expert-based validation by legal professionals in future evaluations would help ensure that assessment metrics reflect not only technical accuracy but also legal soundness and institutional reliability.

Future development will focus on implementing a dynamic feedback loop to continuously refine embeddings and prompts based on user interactions. The system will also be extended to other structured legal domains, such as tax or labor codes, and adapted for deployment in mobile or offline environments to reach underserved populations. Additionally, incorporating validation panels with both legal experts and end users will contribute to building trust, transparency, and broader adoption.

These efforts aim to establish a robust framework for improving public access to legal knowledge, transforming complex regulatory frameworks into comprehensible and actionable information for all citizens.

References

1. C d Perú. Ley n 29571, código de protección y defensa del consumidor. Ley N, 29571. https://www.gob.pe/institucion/indecopi/normas-legales/1244218-29571.
2. Jumpa, A . P.: Barreras de acceso a la justicia, y la justicia comunal como alternativa en el perú. Derecho Soc. **38**, 360–368 (2012)
3. Maximiliano Arango Grajales: La causa jurídica de las cláusulas abusivas. Estudios socio-jurídicos **18**(1), 243–266 (2016)
4. INDECOPI. Informe de resultados de la encuesta nacional de protección del consumidor (2022). https://www.gob.pe/institucion/indecopi/informes-publicaciones/3292219-encuesta-nacional-de-proteccion-del-consumidor-2022

5. de Oliveira Lima, J. A.: Unlocking legal knowledge with multi-layered embedding-based retrieval. arXiv e-prints, pp. 2411, 2024
6. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NPL tasks (2021). https://arxiv.org/abs/2005.11401
7. Xiong, W., et al.: Answering complex open-domain questions with multi-hop dense retrieval (2021). https://arxiv.org/abs/2009.12756
8. Tam, D., et al.: Evaluating the factual consistency of large language models through news summarization (2022). arXiv preprint arXiv:2211.08412
9. de Oliveira Lima, J. A.: Poly-vector retrieval: reference and content embeddings for legal documents (2025). https://arxiv.org/abs/2504.10508
10. Barron, R. C., Eren, M.E., Serafimova, O.M., Matuszek, C., Alexandrov, B.S.: Bridging legal knowledge and ai: retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization (2025). https://arxiv.org/abs/2502.20364
11. Sugathadasa, K., et al.: Legal document retrieval using document vector embeddings and deep learning (2018). https://arxiv.org/abs/1805.10685
12. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-BERT: the muppets straight out of law school (2020). arXiv preprint arXiv:2010.02559
13. Peng, X., Chen, L.: Athena: retrieval-augmented legal judgment prediction with large language models (2024). https://arxiv.org/abs/2410.11195
14. Guo, J., et al.: A deep look into neural ranking models for information retrieval (2019). https://arxiv.org/abs/1903.06902
15. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering (2020). https://arxiv.org/abs/2004.04906
16. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese bert-networks (2019). https://arxiv.org/abs/1908.10084
17. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus (2017). https://arxiv.org/abs/1702.08734
18. Ho, X., Huang, J., Boudin, F., Aizawa, A.: LLM-as-a-judge: Reassessing the performance of LLMs in extractive qa (2025). https://arxiv.org/abs/2504.11972

Faithfulness and Relevance in Stable Normative Domains: A Comparative Analysis of CAG vs. RAG

Elmer A. Collanqui[1,2(✉)], Maribel E. Coaguila[1], Anibal Flores[1], and Honorio Apaza[1,2]

[1] Universidad Nacional de Moquegua, Moquegua, Peru
{mcoaguilam,afloresg}@unam.edu.pe
[2] AiMara Lab, Ilo, Peru
{ecollanqui,hapaza}@aimaralab.com
https://www.aimaralab.com/

Abstract. In higher education, efficient applicant support is crucial but often limited by manual processes and the inherent constraints of large language models (LLMs) regarding factual accuracy within specialized domains. This study conducts a rigorous comparative analysis of Retrieval-Augmented Generation (RAG) and Cache-Augmented Generation (CAG) architectures to determine the optimal approach for an AI-driven chatbot in the university admissions domain. Using the university's 2025 admission regulation as the knowledge base, we evaluated state-of-the-art LLMs, including Google's Gemini (2.5 Flash & 2.5 Pro) and Meta's Llama 3.1 (8B & 70B). Performance was assessed on an expanded dataset of 134 questions using the RAGAs framework, focusing on key metrics like faithfulness, answer relevancy, and response latency. Our results reveal that the CAG architecture generally achieves superior factual accuracy. Notably, the CAG + Gemini 2.5 Flash combination emerged as the optimal configuration in terms of quality, delivering the highest faithfulness (0.970) and answer relevancy (0.757). Crucially, the RAG configuration of this same model registered the lowest response latency, highlighting a clear trade-off between maximal quality and speed. This research provides an evidence-based blueprint for developing highly accurate and efficient AI chatbots for normative domains, demonstrating that a well-designed architecture with a cost-effective model can deliver state-of-the-art performance.

Keywords: RAG · CAG · LLM · Gemini · Llama 3.1 · RAGAS · Chatbots · NLP · Admission · University

1 Introduction

Efficiency in student communication and support is a fundamental pillar of higher education, particularly in sensitive processes such as university admis-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 137–153, 2026.
https://doi.org/10.1007/978-3-032-20322-9_10

sions, where clarity and speed in responses largely shape the applicant's experience. However, many universities still rely heavily on administrative staff to handle an increasing volume of inquiries, leading to delays, inconsistent responses, and a significant operational burden.

In this context, Artificial Intelligence (AI)-powered chatbots particularly those built on Large Language Models (LLMs) have emerged as a viable and scalable solution. Their ability to generate fluent text and sustain natural interactions has been widely documented [9]. Nevertheless, their static knowledge and tendency to produce incorrect or hallucinated responses limit their applicability in domains where factual precision is critical, such as the interpretation of university regulations [7].

To mitigate these limitations, two complementary architectural approaches have emerged: Retrieval-Augmented Generation (RAG) and Cache-Augmented Generation (CAG). RAG enables models to access specific and up-to-date information from external documents in real time, effectively combining parametric and non-parametric memory [13,14]. However, its effectiveness depends on factors such as the semantic quality of the retrieved fragments, contextual coherence, and the efficiency of the retrieval pipeline. It also introduces new challenges in terms of latency, computational cost, and security [8,12,15].

On the other hand, CAG offers an optimized alternative by preloading highly relevant and stable context into memory, allowing for consistent responses with lower latency and without relying on dynamic retrieval mechanisms. This approach is particularly attractive in scenarios with a well-defined information domain, such as institutional regulations.

This study addresses this issue in the context of the National University of Moquegua (UNAM). Rather than proposing a single solution, we conduct a direct comparative evaluation of the two cutting-edge architectures RAG and CAG to empirically determine which of the two offers a more robust solution for answering questions related to the 2025 General Admissions Regulation.

The comparative evaluation was conducted using a corpus of 134 real questions, leveraging state-of-the-art LLMs (Google's Gemini 2.5 Flash & Pro, and Meta's Llama 3.1 in 8B and 70B versions), and metrics from the RAGAs framework [10], focusing on factual faithfulness and answer relevancy. This research contributes both quantitative and qualitative evidence regarding which architecture is more robust for conversational applications in regulatory domains, thereby laying a solid foundation for future implementations in the education sector.

2 Research Objectives

The main objective of this study is to determine the most effective and reliable language generation architecture RAG or CAG for a question-answering chatbot operating in the regulatory domain of university admissions.

To achieve this goal, the following specific objectives are defined:

- **Develop an optimized knowledge base:** Process and structure the 2025 Admissions Regulation of UNAM into a digital corpus using a semantic segmentation strategy by article to maximize contextual coherence.

- **Implement independent evaluation prototypes:** Build two separate functional systems, one for each architecture (RAG and CAG), to serve as the foundation for a direct comparative evaluation.
- **Quantitatively evaluate performance:** Rigorously measure and compare the performance of both architectures using a set of state-of-the-art LLMs (Gemini and Llama 3.1). The evaluation will be conducted on a corpus of 134 real questions using the RAGAs framework to assess factual faithfulness and answer relevancy.
- **Establish a performance benchmark for LLMs:** Use the evaluation results to create a comparative benchmark of the performance of the Gemini and Llama 3.1 models in Q&A tasks involving regulatory documents in Spanish, contributing valuable data to the community.
- **Formulate an evidence-based recommendation:** Analyze the trade-offs between answer faithfulness, relevancy, and computational efficiency for each configuration in order to provide a clear recommendation regarding the optimal architecture and model for the future implementation of an admissions chatbot.

3 Related Work

The advancement of large-scale language models has driven the development of hybrid architectures aimed at improving automated response capabilities when dealing with lengthy and structured documents. In this context, augmented generation approaches such as RAG and CAG stand out as complementary solutions that address the challenges associated with contextualized access to information and efficient text processing [2,3].

This section presents the most relevant prior works that support the comparison between both approaches, with a particular focus on their application to automatic response tasks involving regulatory documents, such as the university admissions regulation used in this study.

3.1 Retrieval-Augmented Generation (RAG)

The RAG model has become established as an effective strategy for integrating external information during the generation process, enabling language models to access specific documents in real time via semantic retrieval. This paradigm has proven especially useful in domains where accurate and up-to-date knowledge is essential.

In this regard, [6] provide a conceptual review of RAG in organizational contexts, highlighting its modular structure, vector retrieval methods, and applicability to document-centric tasks. Their analysis emphasizes how RAG expands the generation context by dynamically incorporating relevant fragments an essential feature for regulatory documents such as institutional guidelines.

From an applied perspective, [5] developed the FolkRAG system, oriented toward cultural heritage collections. This system implements a RAG-based architecture with mechanisms for retrieval, re-ranking, and controlled generation. Their approach demonstrates how the model can be adapted to structured domains by employing specific rules for fragment extraction and contextualization an approach that is clearly transferable to the automated interpretation of academic regulations.

In the study by [4], the authors introduce Proximity, an efficient extension of the RAG model that incorporates an approximate caching system to accelerate access to previously retrieved documents. Their evaluation, conducted in answer generation settings, reports a 59% reduction in latency without compromising semantic accuracy highlighting the value of intermediate storage for recurrent queries in educational environments.

Meanwhile, the study by [2] presents Cache Craft, an architecture that extends the RAG model by incorporating a fragment-level semantic caching system. This mechanism enables the storage and reuse of previously retrieved contexts for similar queries, significantly reducing redundancy in information retrieval and improving computational efficiency. Although the model still relies on a vector database and maintains RAG's retrieval logic, the integration of contextual caching brings it functionally closer to the CAG paradigm particularly in environments where the corpus is static and queries are repetitive.

3.2 Cache-Augmented Generation

The Cache-Augmented Generation (CAG) paradigm emerges as a natural evolution of the traditional Retrieval-Augmented Generation (RAG) framework. Unlike RAG, which dynamically retrieves external information from a database or document corpus at query time, CAG replaces this retrieval phase with advanced caching mechanisms that directly inject pre-stored knowledge into the model's context window. By leveraging cache-based architectures, CAG minimizes the dependency on external retrieval modules and significantly reduces latency and computational overhead, especially in scenarios where the information domain is stable, limited, and frequently reused—such as regulatory documents or academic guidelines.

In a foundational study, [15] introduced a pure CAG approach that eliminates real-time retrieval by preloading the entire knowledge corpus into the model's extended context window. This is achieved through the efficient use of Key–Value (KV) cache mechanisms, which allow the model to reuse hidden states from previous inferences, effectively maintaining relevant information across multiple generations. Their experiments on question-answering benchmarks such as SQuAD and HotPotQA showed that CAG outperformed traditional RAG systems in both semantic accuracy (BERTScore) and response latency, highlighting the efficiency of caching in stable knowledge environments.

Expanding on this line of research, [2] proposed an enhanced CAG architecture to address its scalability constraints through a mechanism known as Adaptive Contextual Compression (ACC). This technique dynamically optimizes the

use of the model's context window by applying hierarchical compression and content pruning, selecting the most relevant knowledge segments via reinforcement learning-based policies. Moreover, the authors introduced a hybrid CAG–RAG configuration that selectively activates retrieval only when the cached information is insufficient. This hybrid approach preserves low latency while ensuring comprehensive knowledge coverage. Evaluations on datasets like HotPotQA and NaturalQuestions demonstrated that ACC–CAG achieved notable improvements in both generation quality (BERTScore) and computational efficiency when compared to conventional RAG and static CAG systems.

3.3 Evaluation of Augmented Generation Systems

Evaluating the quality of LLM-generated text is a complex and multidimensional challenge. Traditional metrics such as BLEU and ROUGE, originally designed for machine translation and text summarization, primarily rely on lexical overlap and therefore fail to capture deeper aspects of meaning, such as semantic coherence, factual faithfulness, and contextual relevance in question-answering (Q&A) systems. As a result, the research community has increasingly emphasized the need for holistic evaluation frameworks that consider the interpretability and factual alignment of generated responses.

In this context, the **RAGAs (Retrieval-Augmented Generation Assessment)** framework has emerged as a robust and widely adopted standard for the evaluation of RAG architectures [10]. Unlike traditional overlap-based metrics, RAGAs employs a *reference-free* evaluation approach, leveraging a large language model (LLM) as an automatic evaluator. This enables a more semantically aware assessment across multiple quality dimensions:

- **Faithfulness**: measures the factual consistency between the generated response and the retrieved context, penalizing hallucinated or unsupported claims.
- **Answer Relevancy**: quantifies how effectively the response addresses the user query in terms of semantic coverage and intent alignment.
- **Context Precision**: evaluates the proportion of retrieved context that is actually relevant to generating the answer.
- **Context Recall**: measures the degree to which the retrieved context contains all the necessary information to answer the query.

In our experiments, we implemented RAGAs using the open-source Python framework ragas. Each metric was computed independently for every question–answer pair in our dataset, with **Gemini 2.5 Flashlite-preview-06–17** serving as the LLM evaluator due to its robust reasoning and factual verification capabilities. For each configuration (RAG and CAG across different models), RAGAs was executed under identical conditions to ensure comparability. The resulting scores were averaged across the full dataset of 134 question–answer pairs, providing a quantitative basis for comparing the relative performance of each architecture.

By adopting RAGAs, our evaluation transcends the limitations of lexical similarity metrics, allowing a more nuanced understanding of the model's ability to retrieve relevant context and generate grounded, semantically coherent responses aligned with user expectations.

3.4 Recent Advances in Large Language Models

The performance of both RAG and CAG architectures is intrinsically linked to the capabilities of the underlying LLMs. The current state of the art is dominated by two main trends: the development of powerful open-source models and the creation of proprietary models with massive context windows. To ensure a representative evaluation, our study selected leading models from both categories.

- **Llama 3.1 Family (Meta):** Representing the forefront of open-source models, Llama 3.1 (8B and 70B) stands out for its optimized architecture and robust performance in reasoning and instruction-following tasks. The availability of different sizes allows for the analysis of the trade-off between efficiency and capability within a single model family.
- **Gemini 2.5 Family (Google):** Representing the frontier of proprietary models, Gemini 2.5 (Flash and Pro) is characterized by large-scale efficiency and an extended context window of up to one million tokens. This capability is particularly relevant for the CAG architecture, which seeks to leverage extended contexts. The Flash variant is optimized for speed, while Pro is focused on maximizing reasoning quality.

The key features of these models are summarized in Table 1. This selection not only enables a comparison between RAG and CAG, but also allows us to examine how these architectures interact with different LLM design philosophies.

Table 1. Key characteristics of the LLMs used in this study.

Model	Parameters	Max Context	Release
Llama 3.1 8B	8B	128K tokens	July 2024
Llama 3.1 70B	70B	128K tokens	July 2024
Gemini 2.5 Flash	ND	1M tokens	June 2025
Gemini 2.5 Pro	ND	1M tokens	June 2025

Note: ND indicates that the number of parameters has not been disclosed.

4 Experimental Design and Evaluated Architectures

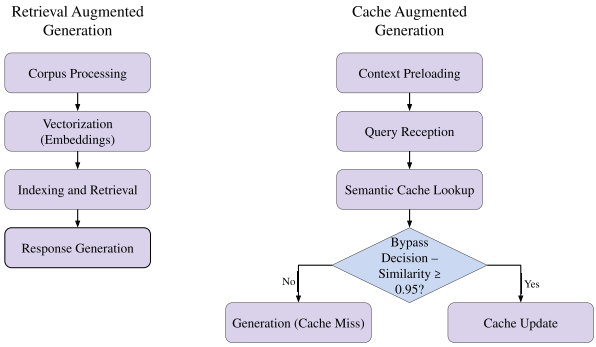


Fig. 1. Comparative workflow between the evaluated architectures. The RAG architecture (left) performs semantic retrieval by article and generates responses based on the most relevant fragment. In contrast, the CAG architecture (right) preloads the entire regulation into the LLM and uses a semantic cache to avoid redundant generation. Both share the same corpus and language models.

To determine the most effective architecture for a university admissions chatbot, a direct comparative evaluation experiment was designed between two independent systems: one based on RAG and the other on CAG. Both implementations use the 2025 Admissions Regulation of UNAM as their sole knowledge source and were evaluated using the same set of LLMs (Gemini and Llama 3.1). The detailed workflow of each architecture is illustrated in Fig. 1.

4.1 Architecture 1: Retrieval-Augmented Generation (RAG)

The RAG pipeline was implemented with a focus on optimizing retrieval quality to ensure a fair comparison.

1. **Corpus Processing:** The regulation text was processed using a **semantic segmentation** strategy, where each chunk corresponds to a complete article of the document. This method ensures the contextual integrity of each regulation is preserved.
2. **Vectorization (Embeddings):** The text fragments were vectorized using the model `sentence-transformers/paraphrase-multilingual-mpnet-base-v2`, chosen for its high performance on semantic tasks in Spanish and its availability via Hugging Face.
3. **Indexing and Retrieval:** The vectors were indexed in a vector database. For RAG retrieval, we used a Chroma vector database with cosine similarity as the distance metric. During inference, upon receiving a user question, the system retrieves only the most relevant article (**top_k=1**). The value of

$k = 1$ was selected to minimize contextual noise and force the LLM to base its response on the most specific and pertinent regulation—a key consideration in legal or normative domains.
4. **Response Generation:** The prompt, which contains the user's question and the retrieved article, is sent to the LLM to generate the final response.

4.2 Architecture 2: Cache-Augmented Generation (CAG)

The CAG architecture was designed to maximize faithfulness and efficiency through a smart caching mechanism that reduces LLM invocations.

1. **Context Preloading:** Upon startup, the system loads the full text of the Admissions Regulation into the LLM's memory or context window.
2. **Query Reception:** The system receives a new user question.
3. **Semantic Cache Lookup:** Before invoking the LLM, the system vectorizes the incoming question and searches for the closest match in a cache database storing pairs of {question, response} from previous interactions.
4. **Bypass Decision:** If the **cosine similarity** between the new question and a cached one exceeds a predefined threshold of **0.95**, the system returns the stored response directly, **bypassing the LLM**. This high threshold ensures that only nearly identical queries trigger the cache.
5. **Generation (Cache Miss):** If no match is found in the cache (a "cache miss"), the question is sent to the LLM, which generates a response using the full context of the preloaded regulation.
6. **Cache Update:** The new question and its freshly generated response are vectorized and stored in the cache database, enriching the system for future queries.

To manage the "cold start" scenario, the study evaluated performance beginning with an empty cache. It is acknowledged that for production deployments, a pre-population phase with known frequently asked questions could improve initial performance. Both architectures were ultimately evaluated using the RAGAs framework on a set of 134 questions.

4.3 Evaluation Methodology

To conduct a quantitative and multifaceted evaluation of the performance of both architectures, a rigorous protocol was designed and executed on a proprietary corpus of 134 question–answer pairs, formulated from real queries related to the Admissions Regulation.

The dataset employed in this study comprises **134 question–answer (Q–A) pairs** derived from the official *Admissions Regulation* of the university. Each question was designed to reflect realistic user inquiries about the admission pro-

cess, while each answer was carefully constructed based on the corresponding article to ensure factual accuracy and consistency. On average, the questions contain **21 tokens**, and the answers **48 tokens**. The dataset is **proprietary** and not publicly available; however, representative examples are provided in Table 2.

Table 2. Representative examples from the Admissions Q–A dataset.

Question	Answer
¿Cuántas veces al año se hace el proceso de admisión?	La Universidad Nacional de Moquegua realiza su proceso de admisión dos veces al año, conforme a su calendario académico anual aprobado por el consejo universitario.
¿se realizara el examen de admisión de manera remota?	No, el examen de admisión no se realizará de manera remota, es presencial.

Response quality was assessed using the state-of-the-art **RAGAs** framework [10]. The analysis focused on two main metrics: **faithfulness**, which evaluates the factual adherence of the response to the provided context; and **answer relevancy**, which measures whether the response directly addresses the user's intent. Additionally, for the RAG architecture, the quality of the retrieval component was assessed using the **context precision** and **context recall** metrics. To ensure reliability in this process, the judging role within the RAGAs framework was carried out by the **Gemini 2.5 Flashlite-preview-06–17** model.

Finally, to evaluate computational efficiency, the **end-to-end latency** of each query was measured, recording the elapsed time in seconds from prompt submission to complete response reception.

5 Results

This section presents the quantitative results of the comparative evaluation between the RAG and CAG architectures. We detail the findings obtained for response quality metrics, evaluated using the RAGAs framework, as well as response time efficiency (latency). Full results are summarized in Table 3 and the following figures.

Table 3. Comparative results on response quality using RAGAs metrics.

LLM Model	Architecture	Faithfulness	Answer Relevancy	Context Precision	Context Recall
Gemini 2.5 Flash	CAG	**0.970**	**0.757**	N/A	N/A
	RAG	0.938	0.666	0.692	0.417
Gemini 2.5 Pro	CAG	0.969	0.687	N/A	N/A
	RAG	0.944	0.651	0.692	0.417
Llama 3.1 8B	CAG	0.913	0.677	N/A	N/A
	RAG	0.895	0.711	0.692	0.417
Llama 3.1 70B	CAG	0.933	0.378	N/A	N/A
	RAG	0.942	0.629	0.692	0.417

Note: Context metrics do not apply (N/A) to the CAG architecture because it does not perform selective retrieval.

5.1 Factual Faithfulness

Factual adherence to the source document is critical in regulatory contexts. As shown in Fig. 2, **CAG generally outperformed RAG** across most models, demonstrating stronger factual alignment with the reference context. However, for the **Llama 3.1 70B** model, **RAG achieved a slightly higher score**, suggesting that the retrieval step benefited this larger model. Overall, **CAG + Gemini 2.5 Flash** achieved the highest score of 0.970, reflecting near-perfect adherence to the regulation.

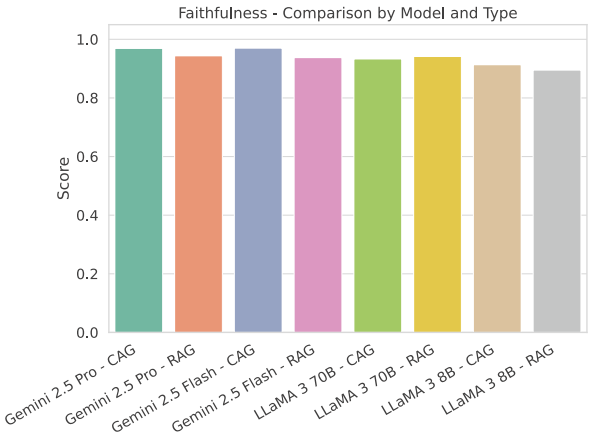


Fig. 2. Faithfulness – Comparison by model and architecture (CAG vs. RAG).

5.2 Answer Relevancy

This metric evaluates whether the generated response directly addresses the user's question. As illustrated in Fig. 3, **CAG + Gemini 2.5 Flash** achieved the highest score (0.757), followed by **RAG + Llama 3.1 8B** (0.711). In contrast, **CAG + Llama 3.1 70B** recorded the lowest performance (0.378), likely due to the model's difficulty in focusing on a full-length corpus.

5.3 Context Retrieval (RAG Only)

The `context_precision` and `context_recall` metrics assess the performance of the retriever in the RAG pipeline. As shown in Figs. 4 and 5, the values remained constant (0.692 and 0.417, respectively), as the same retriever was used across all experiments.

5.4 Response Time Efficiency (Latency)

In addition to quality, the average latency per query was measured for each configuration a critical factor in user experience. The results are presented in Table 4.

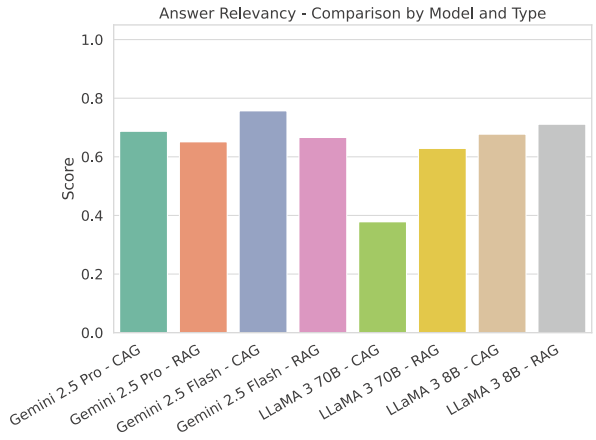


Fig. 3. Answer Relevancy – Comparison by model and architecture (CAG vs. RAG).

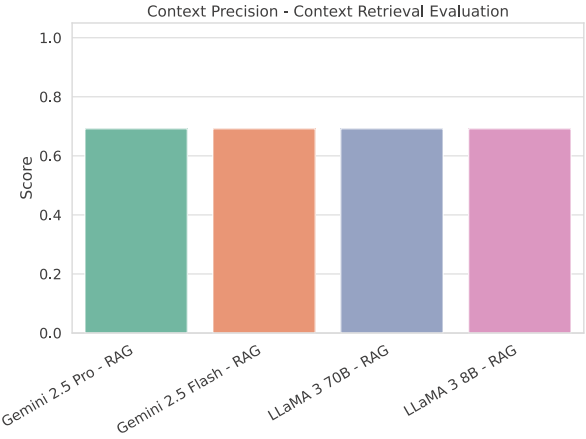


Fig. 4. Context Precision – Retriever evaluation in the RAG architecture.

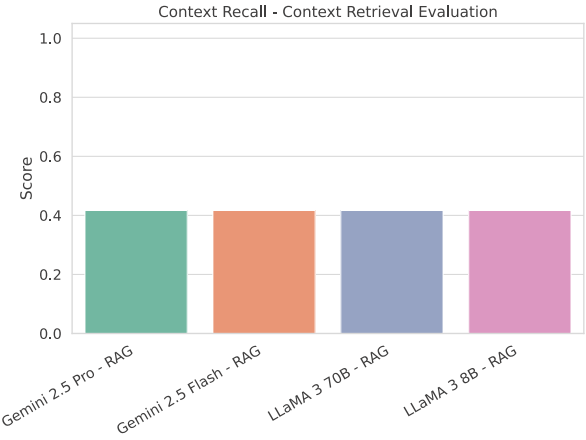


Fig. 5. Context Recall – Retriever evaluation in the RAG architecture.

Latency analysis is crucial for understanding the trade-off between quality and speed. Configurations using Gemini 2.5 Flash were significantly faster, with RAG achieving the lowest latency in the study (0.6052 s). At the opposite end, Gemini 2.5 Pro exhibited the highest latency in both architectures, exceeding 10 s per query. These quantitative results on the balance between response quality and temporal efficiency are further analyzed in the Discussion section.

Although only the end-to-end latency was recorded, the total response time can be conceptually decomposed into three main stages: (1) **retrieval time**, corresponding to vector search and context assembly in the RAG pipeline; (2) **generation time**, dominated by the LLM's text generation process; and (3) **communication overhead**, which includes API and network transfer delays. Based on observed system behavior, generation time was the most influential

Table 4. Average response latency per query (in seconds).

LLM Model	Architecture	Avg. Latency (s)
Gemini 2.5 Flash	CAG	0.7618
	RAG	0.6052
Gemini 2.5 Pro	CAG	10.3859
	RAG	10.1051
Llama 3.1 8B	CAG	1.1682
	RAG	1.4447
Llama 3.1 70B	CAG	2.7420
	RAG	2.2999

component, while retrieval and communication contributed minimally to the overall latency.

6 Discussion

The results presented reveal a clear performance hierarchy among the evaluated architectures and models, allowing us to answer the central question of this study: which is the most robust and reliable approach for a chatbot in a regulatory domain? This section interprets the findings, analyzes their implications, and discusses the study's limitations.

6.1 CAG Superiority in Factual Faithfulness

The most consistent finding of the study is the dominance of the CAG architecture in the *faithfulness* metric (Fig. 2). Across all models, CAG outperformed RAG, reaching a peak score of 0.970 with Gemini 2.5 Flash. This advantage is mainly due to the structural difference between both approaches. RAG depends on a retrieval phase that extracts document fragments based on semantic similarity, whereas CAG processes the entire document directly in the context window.

As shown in Table 3, the RAG configuration achieved a relatively low *Context Recall* (0.417), indicating that it often retrieved incomplete or suboptimal segments of the source text. This limitation arises because, during indexing, the admissions regulation document was divided into small *chunks*. Even though the full document could fit within the model's context window, the retrieval mechanism sometimes selected segments that were semantically close but lacked critical factual details.

By contrast, the CAG architecture avoids this intermediate retrieval step. Since the entire content is provided as input, it eliminates the risk of missing context and achieves consistently higher factual alignment. In compact domains such as this one, where documents are short and self-contained, CAG proves to be both simpler and more accurate. Nevertheless, for larger or multi-document

corpora, RAG remains advantageous due to its scalability and modular retrieval capabilities.

6.2 Answer Relevancy: The Size and Efficiency Dilemma

Unlike faithfulness, the *answer relevancy* metric revealed a more complex and insightful landscape regarding LLM behavior (Fig. 3).

The outstanding performance of **CAG + Gemini 2.5 Flash (0.757)** suggests that, for this task, efficiency and possible "fine-tuning" of a lightweight model may be more critical than raw size. The fact that it outperformed its "Pro" counterpart indicates that a larger model does not necessarily guarantee better alignment with user intent.

Conversely, the most anomalous result was the underperformance of **CAG + Llama 3.1 70B (0.378)**. A possible explanation is the well-known "lost in the middle" phenomenon, where LLMs with very large context windows may lose focus on the original query and instead attend to content in the middle of the context. In contrast, **Llama 3.1 8B** showed much more robust relevancy performance especially in the RAG configuration (0.711) suggesting that its smaller size made it less prone to such "contextual distraction."

6.3 Practical Implications: Trade-Offs Between Quality, Cost, and Latency

The findings of this study have direct implications for the development of AI systems in university environments. The choice of architecture depends not only on response quality, but on a comprehensive analysis of its trade-offs.

First, in terms of **computational cost per token**, the RAG architecture is inherently more efficient. By sending only one retrieved article to the LLM, it minimizes inference cost. Conversely, for every "cache miss," CAG must process the entire regulation, resulting in a significantly higher token cost. This highlights that CAG's economic viability in production hinges critically on achieving a high cache hit rate.

Second, the analysis of **temporal efficiency (latency)**, presented in Table 4, adds another layer of complexity. The data show that **Gemini 2.5 Flash in its RAG configuration was the fastest** (0.6052 s), while its CAG counterpart, despite offering higher quality, was slightly slower (0.7618 s). At the other extreme, Gemini 2.5 Pro was an order of magnitude slower, exceeding 10 s per query, rendering it unsuitable for real-time interactions.

Considering this three-way analysis (quality, cost, latency), the combination of the **CAG architecture with the Gemini 2.5 Flash model remains the top recommendation**. Although its RAG variant is marginally faster, the significant improvements in *faithfulness* (+3.4%) and *answer relevancy* (+13.6%) delivered by CAG justify the small latency increase. The final recommendation is therefore to choose the configuration that ensures maximum quality and reliability while noting that its deployment must include an optimized cache mechanism

to be economically sustainable, and should consider strategies to mitigate the cold start issue.

6.4 Limitations and Future Directions

Despite the methodological rigor, this study has certain limitations that open new avenues for research. First, although the domain of admissions regulations is generalizable, the results are based on a single structured document; future work could evaluate these architectures on more heterogeneous and unstructured corpora. Second, the evaluation is quantitative and automated. Implementing a phase of **human evaluation** could provide more qualitative insight into the naturalness and real-world usefulness of the responses.

Moreover, in the RAG configuration, only a single fragment (*chunk*) was retrieved per query. Future work will analyze the impact of retrieving multiple fragments to assess potential improvements in context recall and overall answer accuracy.

Finally, the main future work direction involves the development of a production-ready system based on the winning configuration. Additionally, the design of an **intelligent hybrid architecture** could be explored, in which a semantic router directs novel or complex queries to the RAG module and recurring questions to the CAG module, thereby aiming to combine the best of both worlds.

7 Conclusion

This study aimed to determine the most effective and reliable language generation architecture RAG or CAG for a chatbot operating in the regulatory domain of university admissions. Through a rigorous comparative evaluation using state-of-the-art LLMs (Gemini and Llama 3.1) and the RAGAs assessment framework, we arrived at clear and meaningful conclusions.

The principal finding of this research is that the **CAG architecture is demonstrably superior to RAG for this use case**. Consistently, CAG achieved higher factual faithfulness in its responses a critical requirement in domains where precision is essential. By operating with the full context, the CAG architecture eliminates the risk of retrieval failures, which we identified as an inherent limitation of the evaluated RAG pipeline.

Additionally, we found that the combination of the **Gemini 2.5 Flash model with the CAG architecture** not only achieved the highest faithfulness score (0.970) but also the highest answer relevancy (0.757), outperforming larger models. This result highlights that a well-designed architecture with an efficient model is more important than the raw size of the LLM. However, our findings also confirm the existence of a **multifactor trade-off**: the superior quality of CAG comes with a higher token cost on cache misses and does not necessarily guarantee the lowest latency, as the fastest configuration (RAG) also used

Gemini 2.5 Flash. This underscores that CAG's viability in production critically depends on the effectiveness of its semantic cache layer.

Ultimately, this work offers an evidence-based recommendation for higher education institutions, positioning the CAG architecture as the preferred approach due to its high reliability. For future work, we propose conducting human evaluations and exploring the design of a hybrid system that adaptively leverages the strengths of both architectures. Beyond the academic domain, our findings suggest that for any domain defined by a canonical and stable knowledge corpus, CAG-style architectures may represent a safer and more reliable implementation paradigm than traditional retrieval-based approaches.

References

1. Agrawal, R., Kumar, H.: Enhancing cache-augmented generation (CAG) with adaptive contextual compression for scalable knowledge integration. arXiv (2025). https://arxiv.org/abs/2505.08261
2. Agarwal, S., et al.: ACM SIGMOD Int. Conf. Management of Data **2025**,(2025). https://doi.org/10.1145/3725273
3. Ravisankar, M., Malladi, R., Pavankumar, R., Arulselvi, K., Konatham, S.: Retrieval-augmented generation classification algorithm for fake-news detection. In: AI for Sustainable Societal Development, CRC Press, pp. 325–341 (2023). https://doi.org/10.1201/9781003529231-25.
4. Bergman, S., et al.: Leveraging approximate caching for faster retrieval augmented generation. In: Proc. EuroMLSys 2025 Conf., (2025). https://doi.org/10.1145/3721146.3721941.
5. Kelly, P., Schild, J., Jafari, A.: FolkRAG: a retrieval augmented generation system for cultural heritage materials. Neural Computing and Applications, vol. (2025). https://doi.org/10.1007/s00521-025-11455-4.
6. Klesel, M., Wittmann, H. F.: Retrieval augmented generation (RAG). Business & Information Systems Engineering, vol. 67, no. 4, pp. 379–384 (2025). https://doi.org/10.1007/s12599-025-00945-3.
7. Huang, Y., Huang, J.: A survey on retrieval-augmented text generation for large language models. arXiv (2024). https://arxiv.org/abs/2404.10981
8. An, B., Zhang, S., Dredze, M.: RAG LLMs are not safer: a safety analysis of retrieval-augmented generation for large language models. In: Proc. 2025 Conf. of the Nations of the Americas Chapter of the Assoc. for Computational Linguistics: Human Language Technologies, Albuquerque, New Mexico, pp. 5444–5474 (2025). https://doi.org/10.18653/v1/2025.naacl-long.281., https://aclanthology.org/2025.naacl-long.281/
9. Zhang, Q., et al.: A survey of graph retrieval-augmented generation for customized large language models. arXiv (2025). https://arxiv.org/abs/2501.13958
10. Es, S., James, J., Espinosa Anke, L., Schockaert, S.: RAGAs: automated evaluation of retrieval augmented generation. In: Proc. 18th Conf. of the European Chapter of the Assoc. for Computational Linguistics: System Demonstrations, St. Julians, Malta, pp. 150–158 (2024). https://aclanthology.org/2024.eacl-demo.16/
11. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: learning to retrieve, generate, and critique through self-reflection. arXiv (2023). https://arxiv.org/abs/2310.11511

12. Merola, C., Singh, J.: Reconstructing context: evaluating advanced chunking strategies for retrieval-augmented generation. arXiv (2025). https://arxiv.org/abs/2504.19754
13. Zheng, X., et al.: Retrieval augmented generation and understanding in vision: a survey and new outlook. arXiv (2025). https://arxiv.org/abs/2503.18016
14. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv (2021). https://arxiv.org/abs/2005.11401
15. Chan, B. J., Chen, C.-T., Cheng, J.-H., Huang, H.-H., Don't do RAG: when cache-augmented generation is all you need for knowledge tasks. In: Proc. Companion Proceedings of the ACM on Web Conf. 2025, pp. 893–897 (2025). https://doi.org/10.1145/3701716.3715490.

BERT-Based Approach for Quechua Dialect Identification

Christian Aranibar Solaligue and Julio Santisteban(✉)

Department of Computing, Universidad Católica San Pablo, Arequipa, Peru
{christian.aranibar,jsantisteban}@ucsp.edu.pe.edu

Abstract. The Identification of dialectal variants poses a significant challenge in indigenous languages due to their complex morphology and the limited availability of annotated data. In the case of Quechua, an agglutinative language with extensive dialectal diversity. This study proposes a Transformer-based approach, specifically using the BERT models, to classify Quechua dialects under low-resource conditions. DeBERTa was selected due to its disentangled attention mechanism and enhanced mask decoder, which enable more effective modeling of contextual relationships and positional information. The methodology involves collecting a multi-dialect Quechua corpus, performing data cleaning and normalization. A stratified cross-validation strategy and hyperparameter optimization using Optuna were implemented to improve model performance. Several architectures, including BERT, DeBERTa, RoBERTa, and ConvBERT, were evaluated under equivalent conditions. Results show that the models achieved improvement on the state of the art with a 0.983 precision and recall in Quechua del Cusco - QUZ with a much larger dataset, demonstrating their efficiency even in linguistically low-resource settings, while ConvBERT has the best performance.

Keywords: Dialect Identification · Quechua · BERT-Based

1 Introduction

Quechua is an indigenous language spoken by 11 million people in southwestern South America. According to Gary Parker (1963), Quechua is a language family with more than 16 variants. Additionally, each region has its variants. Quechua has no formal written form. Quechua is one of the official languages of Peru and Bolivia.

There has been little research, and with poor results, regarding a translator for Quechua, AmericasNLP 2025 Shared Tasks on Machine Translation on its fifth anual avent has atracted few research groups working in machine translator for Quechua - QUY. The top results on the Shared Task for spa-quy with a BLEU of 3,70 and CHRF++ of 32,70 and quy-spa with a BLUE of 12,27 and CHRF++ of 33,02 [4,10].

Although several studies have explored Quechua language processing through hybrid, statistical, and transformer-based models [7,9,15,17], research in this

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 154–166, 2026.
https://doi.org/10.1007/978-3-032-20322-9_11

area remains limited [3]. Let alone an identifier for a Quechua variant. This research is the first work to use a Transformer-based approach to identify variants of the Quechua language.

Quechua is an oral language that uses a written form based on the Latin script, whose orthography is influenced by its Spanish-speaking context. This adaptation allows compatibility with transformer-based language models trained on multilingual datasets. Because of the shared script with languages such as Spanish and English, existing tokenizers can process Quechua without retraining, facilitating fine-tuning and supporting natural language processing efforts for low-resource indigenous languages.

Related Work. Linares et al. in 1997 present a Low-Resourced Peruvian Language Identification Model based on SVM, Multinomial Naive Bayes and other classical machine learning algorithms. We have only one Quechua variant in common, "Quechua del Cusco - QUZ". Linares et al. achieved a precision of 0.94 and a recall of 0.97 in QUZ with a small dataset of 3496 sentences. Our work achieved a precision of 0.983010 and a recall of 0.983009 for QUZ [8].

The adaptation of pre-trained models such as BERT to low-resource languages has been a key issue in natural language processing. [19] proposed an approach to extend *Multilingual BERT (mBERT)* to low-resource languages and dialectal variants not covered in its original training. Their approach is to expand the M-BERT vocabulary by incorporating tokens from the new language, followed by targeted retraining with monolingual data from the target language.

Likewise, [18] proposed *FinBERT*, a monolingual BERT model specifically trained for Finnish, demonstrating that specialised linguistic approaches outperform *mBERT* on low-resource languages and deprioritised variants in multilingual training. Its approach is based on generating adapted vocabulary using BPE, optimising tokenisation to reduce the morphological fragmentation typical of mBERT.

Adapt BERT in low-resource scenarios, combining active learning and selective layer freezing [6]. They used BALD (Bayesian Active Learning by Disagreement) along with MC-Dropout to automatically select the most informative data for labelling, thereby maximising the model's knowledge gain. By freezing the final layers of BERT, they reduced the number of trainable parameters, thereby improving training stability without sacrificing accuracy.

Dialect Identification with BERT. The study [12] proposes a BERT-based multi-task model to identify Arabic dialects at the country and province level. Their approach leverages BERT's bidirectionality to capture subtle contextual differences, achieving high accuracy in dialect classification. To achieve this, they utilised *MarBERT*, a BERT model pre-trained on millions of Arabic tweets.

Similarly, [11] proposed a BERT-based system to identify Arabic dialects in tweets, highlighting its effectiveness in resource-poor languages. Using the multilingual model (mBERT), pre-trained on a large amount of Arabic text, they employed a two-stage approach. First, they adapted the model using millions of

unlabeled tweets, and then performed fine-tuning for classification using labelled tweets.

The authors [2] investigated Arabic dialect identification in resource-poor settings using both traditional models and transformers, highlighting the performance of AraBERT, an Arabic-tuned version of BERT. Despite the challenges presented by data imbalance and similarities between regional dialects. They employed a fine-tuning approach on AraBERT, a pre-trained model on large corpora, and adapted it to a dataset composed of both labelled and unlabeled tweets.

In [20] the used Ro-BERT, a BERT-based model tailored explicitly to the Romanian language, to tackle the task of identifying dialects in resource-limited contexts. Thanks to its pre-training on large Romanian corpora, Ro-BERT demonstrated strong performance compared to multilingual models, distinguishing itself by its ability to capture linguistic nuances specific to the language.

Applications of Transformers in Indigenous Languages. First, [13] developed *IndT5*, a Transformer model based on *T5*, to address the identification and translation of indigenous languages with limited resources and dialectal variants. They utilised a multilingual corpus (IndCorpus), which encompasses several indigenous languages.

In [14] the authors proposed an innovative BERT-based approach to revitalize Inuktitut, a polysynthetic indigenous language with dialectal variants and few resources. They used a BERT-like model, pre-trained with the Masked Language Modeling (MLM) method on monolingual Inuktitut texts, to generate contextual representations, which were then integrated into a Transformer-based neural machine translation. They also incorporated a bidirectional morphological segmenter (bi-LSTM) to handle the complexity of the language's morphology. This approach reduced the need for large volumes of parallel data by leveraging a relatively small corpus adapted through BPE tokenization.

BERT-based models have proven effective in dialect classification and processing tasks for low-resource languages, they still have significant limitations when applied to indigenous languages such as Quechua. The lack of pre-trained models in this language, the absence of tokenizers suitable for its agglutinative morphology, and the scarcity of labelled corpora pose significant barriers to achieving accurate classification of its variants.

2 Method

Our proposal includes the development of DeBERTa, BERT, RoBERTa and ConvBERT models, which will be trained on a clean and representative corpus of Quechua, as well as a training scheme optimised through cross-validation and hyperparameter tuning. Finally, the models were compared using each model's default tokenizers. A cross-validation will be employed using the *k-fold* scheme, along with a hyperparameter tuning stage, to obtain optimized configurations that maximise model performance.

Model Architecture and Training. For each model, a hyperparameter search will be performed, to optimizing variables such as: *hidden_ layers*, *hidden_ size*, *intermediate_ size*, *attention_ heads*, and *epochs*. For the evaluation, stratified cross-validation with N partitions will be applied. The metrics used will be: *accuracy*, *precision*, *recall*, and confusion matrix.

After completing the cross-validation process and selecting the best hyperparameters, a final training of each model will be performed using the whole training set. Its performance will then be evaluated on a previously reserved, independent test set. This evaluation will enable the generalisation capacity of the models to be estimated under real-world conditions, consolidating the results obtained and facilitating objective comparisons between architectures.

Data Source. Data were collected from various sources containing texts in Quechua, covering both its different linguistic variants and a wide range of content types. Table 1 shows sources of the dataset.

Table 1. Corpus of data used in data collection

Corpus	Source
JW300	[1]
AmericasNLP	[1,16]
Many-to-English	[5]

The variants of Quechua considered in this research correspond to different regions of Peru and Ecuador, according to the ISO classification. 639-3 as shown in Table 2.

Table 2. Variants of Quechua and the size of each corpus [1,5,16].

ISO 639-3 Code	Name	Country	Number of sentences
QUZ	Cusco Quechua	Peru	123 347
QUY	Ayacucho Quechua	Peru	110 397
QVI	Imbabura Highland Quechua	Ecuador	44 733
QUG	Chimborazo Highland Quechua	Ecuador	18 653
QUW	Tena Lowland Quechua	Ecuador	15 380

Data Cleaning. An exhaustive cleaning and normalization process was performed on the collected corpus, aimed at ensuring the quality, consistency, and

linguistic relevance of the data. This process included both basic text transformations and semantic and structural filtering. The application of these transformations enabled us to build a refined, representative corpus. The main tasks are as follows.

1. The data were filtered by ISO code, retaining only the sentences corresponding to recognized variants of Quechua: `quz`, `quy`, `qvi`, `qug`, and `quw`.
2. All characters were converted to lowercase to maintain typographic consistency.
3. Special characters and punctuation marks were removed, preserving only letters, accents, apostrophes, and spaces, in accordance with the orthographic peculiarities of Quechua.
4. Stopwords were removed in both Spanish and English to reduce noise in the analysis, as they are very common words used in sentences and would provide a bias.
5. Non-Quechua words were filtered using a multilingual model, fastText, which allowed for the identification and exclusion of sequences in Spanish and English.
6. Invalid tokens, such as isolated characters, single letters, or irrelevant strings, were removed.
7. Linguistic labels were encoded using one-hot encoding to represent classes for the multiclass classification task adequately.

Hyperparameter Selection and Optimization. For each model, an automated hyperparameter search was performed using the *Optuna* library. This search consisted of running three *trials* per *fold* during the cross-validation process, to optimize the following key variables of the *Transformer*-based architectures:

- **Number of epochs (`epochs`):** Between 4 and 6 epochs were tested, in increments of 2, to balance learning.
- **Hidden layers (`num_hidden_layers`):** Configurations with 4 and 6 layers in the *Transformer* encoder were evaluated.
- **Hidden layer dimension (`hidden_size`):** Sizes of 256, 512, and 768 were explored for the hidden layers.
- **Intermediate layer size (`intermediate_size`):** Dimensions of 256, 512, and 768 were considered, compared to the typical value of 3072, for the encoder's *feed-forward* layer.
- **Attention heads (`num_attention_heads`):** Configurations with 4 and 8 heads per attention layer were tested.

The objective of this optimization was to identify efficient combinations that balanced accuracy and computational complexity. Twelve experiments were performed per model (3 *trials* $\times$ 4 *folds*). In addition, two further tests were conducted per model: one for the final training with the best hyperparameters and another for evaluation on the independent test set, totalling 14 runs per model.

Cross-Validation. Model performance was validated using stratified cross-validation with $k = 4$ partitions. This strategy enabled a balanced distribution of classes within each fold, ensuring that all variants were equally represented during both training and evaluation.

Evaluation with Default Tokenizers. Evaluating the performance of the selected architectures using the default tokenizers for each model, thereby establishing a baseline for comparison.

3 Results

The purpose of the experiments was to evaluate and compare the performance of the selected architectures (BERT, RoBERTa, ConvBERT, and DeBERTa) in the classification task. To achieve a robust evaluation of the models' performance:

Evaluation with default tokenizers. The performance of the four architectures was analyzed using each model's default tokenizers, establishing a baseline for comparison. This stage enabled us to evaluate the models' generalization capacity without specific adaptations to Quechua.

Optimization and Cross-Validation. The best hyperparameters identified for each model are summarized in Table 3.

Table 3. Optimal hyperparameters per model identified after optimization.

Model	Epochs	Layers	Hidden Size	Intermediate Size	Heads
BERT	6	6	768	512	4
DeBERTa	6	6	512	768	4
RoBERTa	6	4	768	768	4
ConvBERT	6	4	768	512	8

To complement this analysis, parallel coordinate and contour plots were generated for each model, using the visualization tools provided by *Optuna*. These plots enable the analysis of how different hyperparameters affect the evaluation metric, facilitating the identification of optimal configurations and relationships between variables as shown on the Figs. 1, 2, 3, 4, 5, 6, 7 and 8.

BERT. For BERT, the graphs indicate that configurations with a higher number of epochs (6), greater depth (6 hidden layers), and a larger hidden layer size (768) tend to yield the best performance. Similarly, an intermediate size of 512 and a moderate number of attention heads (4) appear to contribute to more stable performance. These results suggest that BERT benefits from greater modeling power, with no clear evidence of overfitting within the explored range.

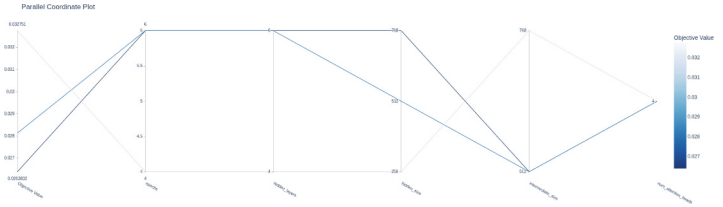


Fig. 1. Parallel coordinates obtained for the BERT model.

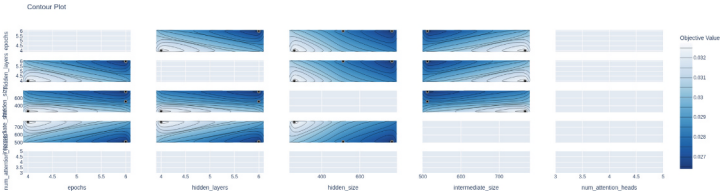


Fig. 2. Contour plot obtained for the BERT model.

DeBERTa. For DeBERTa, the visualizations show greater sensitivity to the size of the hidden and intermediate layers, with better results observed in configurations with high values (768). Increasing the number of epochs (6) also has a positive effect, although with less variability than in BERT. The most consistent configurations are presented with 6 hidden layers and 4 attention heads, suggesting an adequate balance between capacity and generalization.

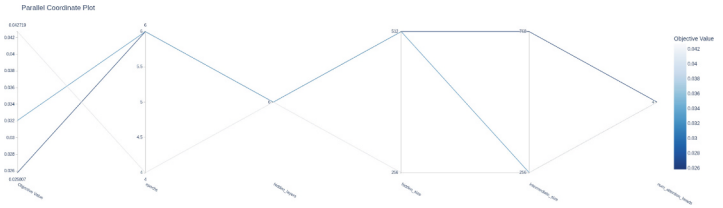


Fig. 3. Parallel coordinates obtained for the DeBERTa model.

RoBERTa. In the case of RoBERTa, the graphs indicate that the best configurations are achieved with more compact architectures, specifically a smaller number of hidden layers (4) combined with high hidden and intermediate sizes (768). This combination appears to favour performance, possibly due to its greater representation capacity without the need for a deep architecture. On the other hand, both the number of attention heads and the number of epochs have a less significant impact compared to other models.

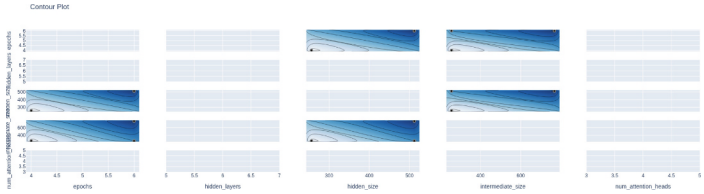


Fig. 4. Contour plot obtained for the DeBERTa model.

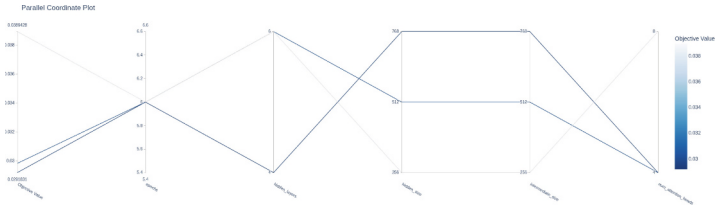


Fig. 5. Parallel coordinates obtained for the RoBERTa model.

ConvBERT In ConvBERT, configurations incorporating a larger number of attention heads (8) and a smaller number of hidden layers (4) perform best. Furthermore, an intermediate hidden size (512) and a reduced intermediate dimension also appear to benefit performance. These results suggest that, due to its hybrid convolution-based architecture, ConvBERT is better able to leverage a larger attention capacity than excessive depth.

3.1 Final Training and Evaluation

After selecting the optimal hyperparameters for each model, the final training and evaluation were conducted on the independent test set. To achieve this, metrics were monitored by epoch throughout the training of each model, considering both the evolution of loss and the performance metrics, including accuracy, precision, and recall(Figs. 9 and 11).

Comparison of metrics by epoch. Figure 10 shows the evolution of the evaluation metrics (accuracy, precision, and recall) during the training of the four models analyzed. All exhibit stable behavior, with progressive improvements over

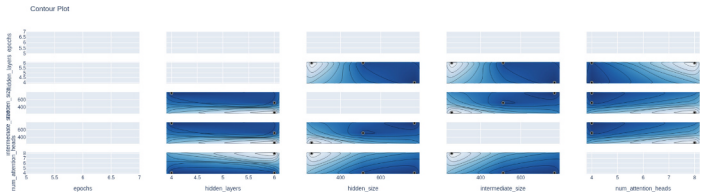


Fig. 6. Contour plot obtained for the RoBERTa model.

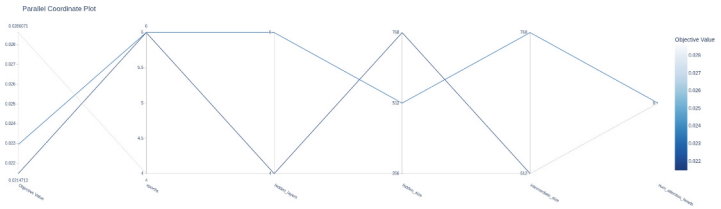


Fig. 7. Parallel coordinates obtained for the ConvBERT model.

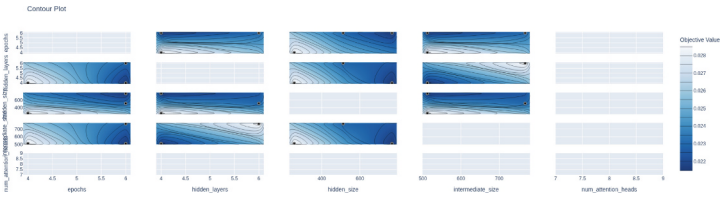


Fig. 8. Contour plot obtained for the ConvBERT model.

the epochs. The differences between models are minimal, suggesting that their performance is more influenced by their particular configurations than by fundamental differences in their architectures.

Table 4. Summary of final metrics for the models evaluated

Model	Training Loss	Validation Loss	Accuracy	Precision	Recall
BERT	0.0120	0.09073	0.978993	0.978992	0.978993
DeBERTa	0.0165	0.08168	0.978817	0.978811	0.978817
RoBERTa	0.0131	0.08758	0.979025	0.979029	0.979025
ConvBERT	0.0094	0.07797	**0.983009**	**0.983010**	**0.983009**

As shown in Table 4, the four models achieved very similar values for their evaluation metrics, with ConvBERT exhibiting slightly better overall performance. However, the differences are minimal, indicating that all approaches are equally valid for the task at hand.

3.2 Confusion Matrix Analysis

The confusion matrices for the four models evaluated confirm highly accurate and consistent performance in all cases. Each model exhibits a notable concentration of correct predictions on the main diagonal, indicating a high success rate per class.

The few incorrect predictions are scattered off the diagonal and mostly correspond to confusions between classes with linguistic similarities. However, these

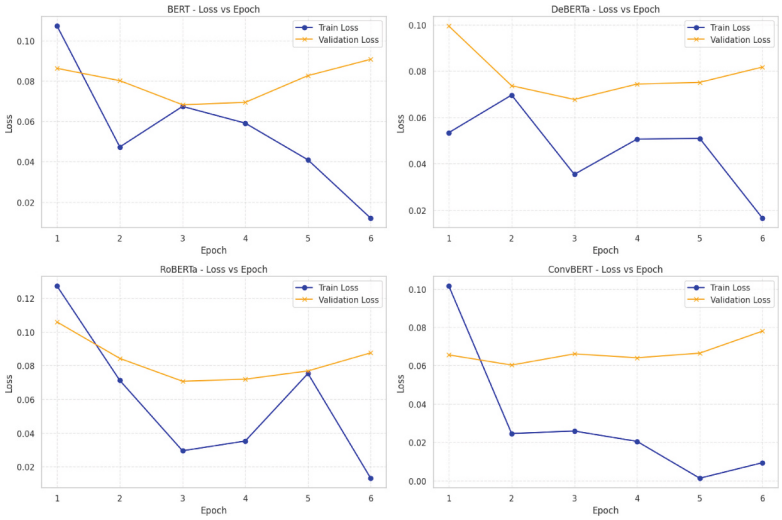


Fig. 9. Evolution of the *Training Loss* and *Validation Loss*.

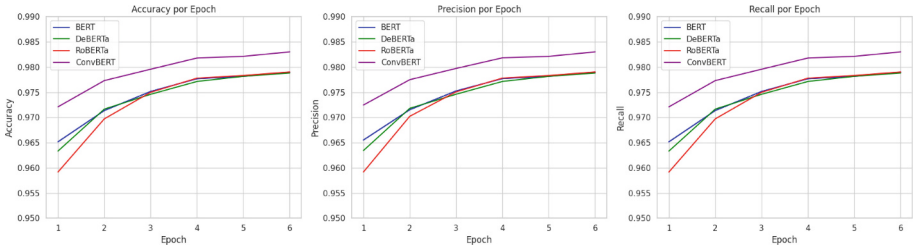
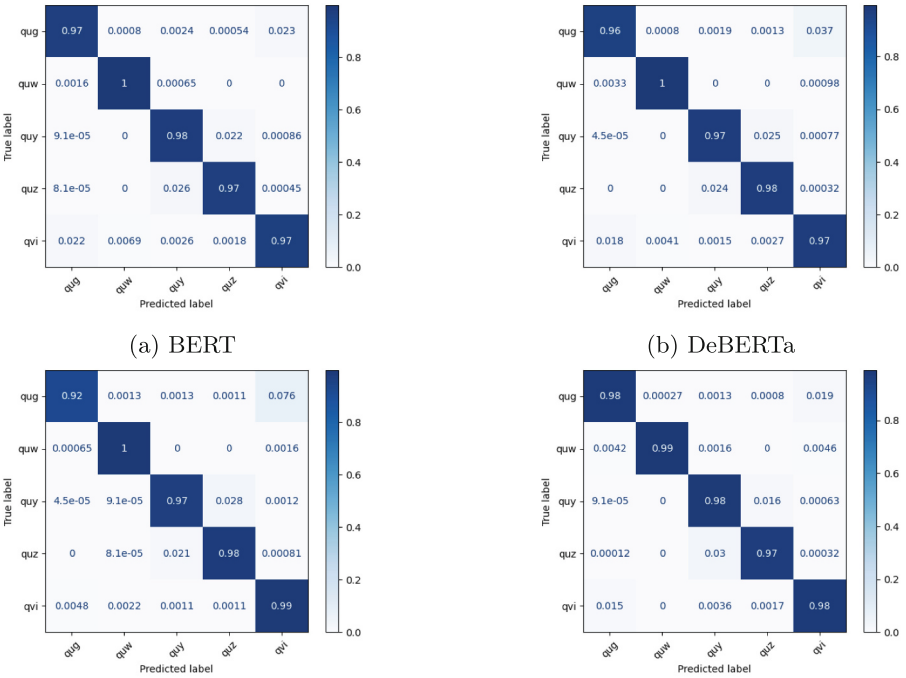


Fig. 10. Evolution of *Accuracy*, *Precision* and *Recall* by epoch evaluated.

deviations are minor and do not substantially impact the overall performance of the models.

It is noteworthy that all models achieve accuracies above 98% for the main classes, reflecting excellent generalization capabilities. The differences between them are minimal, and no systematic error pattern favouring any particular one is identified.

Together, the confusion matrices complement the quantitative results obtained with the aggregated metrics, demonstrating not only high average performance but also reliable and balanced classification across classes.



(a) BERT (b) DeBERTa

(c) RoBERTa (d) ConvBERT

Fig. 11. Confusion matrices in the test set for each model evaluated.

4 Conclusion

This research proposes an approach based on the DeBERTa model for the classification of dialectal variants of Quechua, an indigenous language with complex morphological features and a scarcity of labeled data. Preliminary results show that the use of Transformer models, particularly ConvBERT, has proven effective in classifying Quechua variants, even in a resource-limited context.

The performance of models, including DeBERTa, BERT, RoBERTa, and ConvBERT, was comparable, suggesting that Transformers are robust tools for addressing the challenge of dialect classification, even in scenarios with limited linguistic data. Theoretically, the hyperparameter optimization process, coupled with stratified cross-validation, contributed to significant improvements in the models' accuracy and efficiency, highlighting the usefulness of tools like Optuna for this type of task.

The results obtained support the viability of *Transformer* models for low-resource languages and reinforce the importance of adapting text processing strategies to linguistic specificities, such as the agglutinative morphology of Quechua. This approach not only advances the classification of Quechua dialectal

variants but also sets a precedent for applying similar models to other indigenous languages with similar challenges.

Disclosure of Interests. The authors declare that they have no financial, personal, or professional competing interests that could have appeared to influence the work reported in this paper.

References

1. Agić, Ž., Vulić, I.: JW300: a wide-coverage parallel corpus for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3204–3210. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1310, https://www.aclweb.org/anthology/P19-1310
2. AlShenaifi, N., Azmi, A.: Arabic dialect identification using machine learning and transformer-based models: submission to the NADI 2022 shared task. In: Bouamor, H., et al (eds.) Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pp. 464–467. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (2022). https://doi.org/10.18653/v1/2022.wanlp-1.50, https://aclanthology.org/2022.wanlp-1.50/
3. Asillo Congora, J., Santisteban, J., Lazo Vasquez, R.: UCSP submission to the AmericasNLP 2025 shared task. In: Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pp. 84–91. Association for Computational Linguistics, Albuquerque, New Mexico (2025). https://doi.org/10.18653/v1/2025.americasnlp-1.9, https://aclanthology.org/2025.americasnlp-1.9/
4. De Gibert, O., et al.: Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In: Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pp. 134–152. Association for Computational Linguistics, Albuquerque, New Mexico (2025).https://doi.org/10.18653/v1/2025.americasnlp-1.16, https://aclanthology.org/2025.americasnlp-1.16/
5. Gowda, T., Zhang, Z., Mattmann, C., May, J.: Many-to-english machine translation tools, data, and pretrained models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 306–316. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-demo.37, https://aclanthology.org/2021.acl-demo.37
6. Grießhaber, D., Maucher, J., Vu, N.T.: Fine-tuning BERT for low-resource natural language understanding via active learning (2020). https://arxiv.org/abs/2012.02462
7. Huarcaya Taquiri, D.: Traducción automática neuronal para lengua nativa peruana (2020)
8. Linares, A.E., Oncevay-Marcos, A.: A low-resourced peruvian language identification model. In: CEUR Workshop Proceedings. CEUR-WS, pp. 57–63 (2017)
9. Llitjós, A.F., Aranovich, R., Levin, L.: Building machine translation systems for indigenous languages. In: Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA (2005)

10. Mager, M.,et al(eds).: Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). Association for Computational Linguistics, Albuquerque, New Mexico (2025).https://doi.org/10.18653/v1/2025.americasnlp-1.0, https://aclanthology.org/2025.americasnlp-1.0/
11. Mansour, M., Tohamy, M., Ezzat, Z., Torki, M.: Arabic dialect identification using BERT fine-tuning. In: Zitouni, I., et al. (eds.) Proceedings of the Fifth Arabic Natural Language Processing Workshop. pp. 308–312. Association for Computational Linguistics, Barcelona, Spain (Online) (2020). https://aclanthology.org/2020.wanlp-1.33/
12. Mekki, A.E., Mahdaouy, A.E., Essefar, K., Mamoun, N.E., Berrada, I., Khoumsi, A.: Bert-based multi-task model for country and province level modern standard arabic and dialectal arabic identification (2021). https://arxiv.org/abs/2106.12495
13. Nagoudi, E.M.B., Chen, W.R., Abdul-Mageed, M., Cavusogl, H.: Indt5: a text-to-text transformer for 10 indigenous languages (2021). https://arxiv.org/abs/2104.07483
14. Ngoc Le, T., Sadat, F.: Revitalization of indigenous languages through pre-processing and neural machine translation: the case of inuktitut. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 4661–4666. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020).https://doi.org/10.18653/v1/2020.coling-main.410, https://aclanthology.org/2020.coling-main.410/
15. Rios, A., Göhring, A.: Machine learning disambiguation of quechua verb morphology. Association for Computational Linguistics (2013)
16. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (2012). http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf
17. Vilca, H.D.C., Calderón, V.C.D.M., Mariño, F.C.C., Calderón, E.F.M.: Traductor automático en linea del español a quechua, basado en la plataforma libre y código abierto apertium. Revista de Investigaciones de la Escuela de Posgrado de la UNA PUNO **5**(3), 81–99 (2009)
18. Virtanen, A., et al.: Multilingual is not enough: Bert for finnish (2019). https://arxiv.org/abs/1912.07076
19. Wang, Z., K, K., Mayhew, S., Roth, D.: Extending multilingual BERT to low-resource languages (2020). https://arxiv.org/abs/2004.13640
20. Zaharia, G.E., Avram, A.M., Cercel, D.C., Rebedea, T.: Exploring the power of romanian BERT for dialect identification. In: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 232–241. International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online) (2020). https://aclanthology.org/2020.vardial-1.22/

Quechua Speech Synthesizers

Edwin Huamán Curo(✉) and Ferdinand Pineda Ancco

Pontificia Universidad Católica del Perú (PUCP), Lima, Peru
{a20223941,fpineda}@pucp.edu.pe

Abstract. This research addresses the loss of Quechua's cultural and linguistic heritage, a language classified as endangered according to UNESCO reports. We focused on the Chanca variety of Quechua (Chanca runasimi), spoken in the departments of Andahuaylas, Ayacucho, and Huancavelica, which has approximately one million speakers.

The objective of this paper is the implementation of four Quechua text-to-speech synthesizers: Tacotron 2 + WaveNet, FastPitch + WaveGlow, VITS and XTTS v2; using complementary pre-training and fine-tuning, with a data set of 2,300 high-quality recordings. We performed an objective evaluation of the prosodic and spectral quality with F0 RMSE and SECS, and a subjective evaluation of the naturalness and similarity with MOS-N and MOS-S, respectively. The purpose is to provide open resources that support linguistic revitalization and reduce technological barriers in Quechua-speaking communities.

The results indicate that the end-to-end XTTS v2 model offers greater naturalness and similarity to the original voice. Differences in fundamental frequency of up to 12 Hz on average and SECS values of around 0.10 were also observed, indicating more accurate and consistent models. Furthermore, the MOS-N and MOS-S scores reached 4.9 and 4.8, respectively, demonstrating strong acceptance of the model among the evaluators.

Finally, this approach seeks to contribute to the preservation of the Quechua language through technological resources that promote its use and dissemination in diverse contexts. These synthesizers will facilitate the creation of educational and cultural tools, increase accessibility to the language, and strengthen its relevance in Peru's society.

Keywords: Speech synthesis · Deep learning · TTS · Language preservation · Quechua · Chanca Runasimi

1 Introduction

Quechua is considered an endangered language according to UNESCO's *Atlas of the World's Languages in Danger* [19], and is recognized by the Peruvian state as part of Peru's Intangible Cultural Heritage and as a fundamental linguistic right [17]. However, people whose native language is Quechua face numerous challenges in accessing information and technology in their language due to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 167–183, 2026.
https://doi.org/10.1007/978-3-032-20322-9_12

the limited availability of adapted tools [9]. This not only reinforces technological exclusion, but also perpetuates socioeconomic barriers and stereotypes that associate Andean people with negative perceptions such as laziness or lack of education, affecting their social mobility [25].

UNESCO highlights that while Artificial Intelligence (AI) has transformative potential to address global challenges, it also carries significant risks of exacerbating inequalities and infringing on human rights if ethical aspects are not considered [8]. This is crucial in contexts where historically marginalized communities, such as indigenous peoples, must be included in the design and use of emerging technologies.

In this context, a Quechua speech synthesizer would not only enhance access to information in one's native language but also reduce cultural and technological disparities, while valuing linguistic diversity and fostering more inclusive technology. Furthermore, artificial intelligence has significant potential to preserve indigenous identities, safeguard cultural heritage, and revitalize languages, thus strengthening the synergy between technological advancements and respect for ancestral cultures. [8].

A text-to-speech (TTS) system for Quechua represents a significant opportunity for the preservation and revitalization of this language. This tool has potential in the following applications:

Digitization of Cultural Heritage: It would facilitate the conversion of oral narratives and traditional knowledge into digital formats.

Education: It would allow the development of digital pedagogical resources to support the transmission of the language, especially for children and youth.

Service Accessibility: It would allow the implementation of Quechua voice interfaces for public services and digital platforms, improving access to information in areas such as health, education, and agriculture, particularly in rural regions.

Integration into contemporary digital tools would contribute to the vitality of Quechua in today's digital environment, supporting its continued relevance as a living language.

Despite significant advances in the development of TTS models, there is still a notable gap in TTS systems for low-resource languages [16]. This imbalance is mainly due to: *Data Scarcity:* Most low-resource languages lack the large annotated speech corpora necessary for training robust models. *Resource Limitations:* Limited technological and research resources for Quechua have created a substantial gap in access to digital tools. Beyond accessibility, we outline how TTS supports scalable pedagogical content, digital cultural media, and upstream ASR data augmentation, providing practical levers for Quechua revitalization.

1.1 Previous and Related Work

Text-to-Speech (TTS) has undergone a major transformation in the last decade, driven primarily by advances in deep learning and neural networks [29]. These developments have enabled synthetic speech with high naturalness and expressiveness in many languages worldwide. Tan et al. provide a chronological overview of neural speech synthesis models up to 2021 [29]. Table 1 lists the key TTS models relevant to this study up to February 2024.

Table 1. Key TTS Models and Vocoders Evolution

Date	Model	Type	Features
2016.12	WaveNet	Vocoder	Autoregressive waveform generation
2017.06	Tacotron	Acoustic	Seq2seq with attention
	Tacotron 2	Acoustic	Improved Tacotron with WaveNet
2018.06	TransformerTTS	Acoustic	Transformer-based
2018.12	FastSpeech	Acoustic	Non-autoregressive
	WaveGlow	Vocoder	Flow-based vocoder
	MelGAN	Vocoder	Mel-spectrogram GAN
2020.06	FastSpeech 2	Acoustic	Improved control
	FastPitch	Acoustic	Pitch-controlled FastSpeech
2020.12	Glow-TTS	E2E	Normalizing flows
	HiFi-GAN	Vocoder	High-fidelity GAN
2021.06	VITS	E2E	Variational + adversarial
	Grad-TTS	E2E	Gradient-based
2022.09	Tortoise	E2E	Open source
	Bark	E2E	Multimodal
2023.06	VALL-E	E2E	Large-scale
	XTTS v2	E2E	Zero-shot multilingual
2024.02	StyleTTS 2	E2E	Human-level emotional realism

Technological development in artificial intelligence applied to low-resource languages includes diverse efforts to preserve and revitalize them. Some notable advances are in the following areas:

Virtual Quechua Newscaster: Illariy ("dawn" in Quechua), created in 2023 by the Universidad Nacional Mayor de San Marcos, is the first virtual newscaster to deliver news in Quechua [7]. This initiative underscores the growing interest in applying artificial intelligence to promote and preserve indigenous languages.

Automatic Speech Recognition (ASR): Recent research [26] shows important advances in ASR for indigenous languages such as Quechua, Guaraní, Bribri, Kotiria, and Wa'ikhana. In particular, Quechua achieved a notably low error rate (12.14%), demonstrating the viability of deep learning for minority languages.

Hinantin—Quechua TTS Implementation: The Hinantin project [30] implements a TTS system based on the selection of cluster units using the Festival Speech Synthesis System for Quechua II-C (Cusco). Despite limitations from prosodic complexity and data scarcity, Hinantin represents significant progress in digital inclusion and Quechua language revitalization.

Siminchik—Southern Quechua Speech Corpus (2018): The first large-scale Southern Quechua speech corpus contains **97 h** spontaneous radio conversations across the *Chanca* and *Collao* varieties of Quechua. Transcriptions were crowdsourced from native speakers under a unified orthography, and preliminary ASR and language modeling experiments were reported, establishing a foundational public resource for Quechua speech technology [3].

Huqariq—Multilingual Native Languages Corpus (2022): *Huqariq* is a **multilingual** corpus of Peruvian indigenous languages for **ASR**, **LID**, and **TTS**, totaling **220 h** transcribed by over **500** volunteers. It employs *crowdsourcing* for sustainable collection and targets scaling from 4 to 20 languages. Reported recognition results validate dataset quality, positioning it as the largest speech resource for Peru's native languages [33].

Data Augmentation for Low-Resource Quechua ASR (2022): Zevallos et al. propose a data augmentation pipeline that combines text augmentation and synthetic speech to improve ASR for agglutinative, low-resource languages. Using *wav2letter++*, they train on ∼99 h of original Quechua speech plus ∼99 h of synthetic audio, achieving an 8.73% relative WER reduction to 22.75%. The study demonstrates that carefully generated synthetic data can substantially strengthen Quechua ASR under data scarcity [32].

2 Methodology

A corpus of ≈6 h of Quechua speech was collected and curated, balanced by gender and age range; metadata were validated, *outliers* removed, and the data were split into `train`/`test` sets. The preprocessing included noise reduction up to SNR $\geq$ 20 dB, transcript correction, and normalization of UTF-8, as well as segmentation into clips of $\approx$ 3—20 s for efficient *training*.

For textual input, we used a minimal phonemization frontend based on the Latin American Spanish pronunciation dictionary `es_LA_nv230301.dict` (a resource known in NVIDIA NeMo TTS recipes): if a word was found in the

lexicon, it was replaced by its phonemic (IPA) sequence; otherwise, a Spanish phonemizer (phonemizer/espeak-ng) was applied as a fallback. We did not add Quechua-specific morphophonemic rules: the system operates with Spanish phonemization as-is; the same symbol inventory was kept for training and inference (in VITS we set `use_phonemes=True` with `phoneme_language=es`). Regarding the vocoder, we followed canonical pairings per architecture: WaveNet with Tacotron 2 and WaveGlow with FastPitch; for the end-to-end models VITS and XTTS v2, no external vocoder was used, since they rely on their integrated adversarial/HiFi-GAN-like decoder.

Each architecture (Tacotron 2 + WaveNet, FastPitch + WaveGlow, VITS, XTTS v2) was trained using complementary pre-training or fine-tuning with NVIDIA NeMo/Coqui TTS on RTX GPUs (A5500/4090/A6000), saving the best *checkpoint* per model (≈1000–2000 epochs). Objective evaluation measured prosody and spectral quality using F_0 RMSE and SECS, subjective evaluation assessed naturalness and speaker similarity via MOS-N and MOS-S with $N = 20$ Quechua-speaking raters (Chanca variety) working in education and healthcare. We conducted a web-based listening test using a personalized interface (https://pucpehuaman.github.io/tts_qu/). For each utterance, the stimuli included one natural reference and the synthesized outputs from each model; presentation order was randomized across raters and items. We collected MOS-N (naturalness; 1–5 Absolute Category Rating, ACR) and MOS-S (speaker similarity to the reference; 1–5 ACR), discarding incomplete submissions. Following ITU-T P.800 [10] as feasible in a remote setting—with P.808-style checks—we report per-system means with 95% confidence intervals. Although $N = 20$ limits statistical power and external validity, this constitutes a rigorous first step toward standardized subjective evaluation in Quechua TTS.

2.1 Theoretical Foundations

Text-to-Speech (TTS). Text-to-Speech is the technology that converts written text into spoken voice. Its fundamental objective is to facilitate human–computer communication by enabling textual information to be accessed auditorily [6]. Early TTS systems relied on concatenative methods that stitched together pre-recorded speech fragments [21]. With technological progress, rule-based and statistical models were introduced. In the last decade, deep learning techniques have revolutionized the field, enabling more natural and fluent systems [31].

Speech Signal Processing

Audio Signals and Frequency Analysis. Audio signals are representations of sound waves over time with main properties including frequency (determining pitch) and amplitude (influencing loudness) [31]. Frequency analysis decomposes an audio signal into its frequency components using the Fourier transform [23]. A spectrogram provides a visual representation showing how energy of different frequencies varies over time [11].

Mel Spectrogram. The mel spectrogram uses a frequency scale based on human auditory perception, known as the mel scale [28]. This scale relates physical frequency (Hz) to perceived pitch through:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

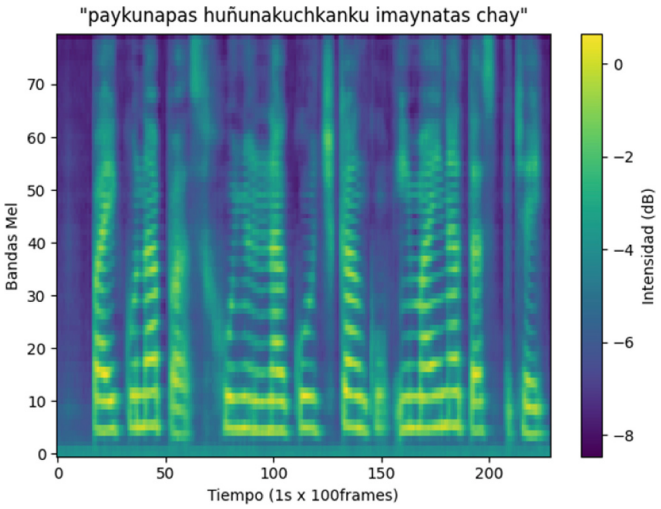


Fig. 1. Mel spectrogram of a Quechua voice.

This transformation provides a better representation of perceptual features, effective reduction in dimensionality, and better alignment with the human auditory system [18].

Evaluation Metrics. To comprehensively evaluate Quechua TTS models, we applied a combined approach that included two objective metrics (F0 RMSE and SECS) focusing on acoustic fidelity, and two subjective metrics (MOS-N and MOS-S) targeting perceptual assessment (Fig. 1).

– **F0 RMSE:** This objective metric compares the fundamental frequency (F0)–the lowest frequency of audio at a given time–between a reference and a synthesized signal [2].

$$\mathrm{RMSE}_{F0} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F0_{\mathrm{pred},i} - F0_{\mathrm{ref},i}\right)^2}. \tag{2}$$

– **SECS (Speaker Embedding Cosine Similarity):** This objective metric evaluates the similarity between two speaker-embedding vectors [14].

$$\text{SECS} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|}, \tag{3}$$

where **A** and **B** are the embedding vectors extracted from synthesized and reference audio, respectively; SECS ranges from −1 (completely different) to 1 (identical).

– **MOS-N (Mean Opinion Score—Naturalness):** This subjective metric measures the naturalness of synthesized speech on a 1–5 scale, where 5 indicates "completely natural" and 1 indicates "totally artificial/robotic" [13].

$$\text{MOS-N} = \frac{1}{N} \sum_{i=1}^{N} \text{score}_i. \tag{4}$$

– **MOS-S (Mean Opinion Score—Similarity):** This subjective metric measures similarity between synthesized speech and a reference voice on a 1–5 scale, where 5 indicates that the voice sounds "exactly the same" as the reference and 1 indicates "completely different" from the original speaker [13].

$$\text{MOS-S} = \frac{1}{N} \sum_{i=1}^{N} \text{score}_i. \tag{5}$$

2.2 Data Preprocessing

Dataset. Thanks to arrangements by Pontificia Universidad Católica del Perú, we accessed a dataset of 9,490 audio files with over 60 h of Quechua speech and the corresponding transcriptions, the Lurin corpus, which contains sentences collected from multiple sources, specifically the Collao–Quechua and Chanka–Quechua dictionaries published by the Peruvian Ministry of Education (2021) [32]. We found that 80% of the audio did not meet the expected quality, with the signal-to-noise ratio below 20 dB, as required in [27]. There was an 80% transcription error rate across multiple Quechua varieties; there were about 500 speakers, roughly 75% male and 25% female. Finally, almost 60% of the audios had imprecise phrase cuts. After cleaning, we obtained 2,300 audios with nearly 6 h of high-quality Quechua speech and correct transcriptions, balanced 50% male/female, and of the Chanca variety. These 2,300 audios served as input to train our models; we used 3–6 h subsets, depending on the model and the available training hardware, with the goal of achieving the best training performance.

Audio Cleaning. The cleaning process was extensive and meticulous. The strategy was to identify the audio (by gender, variety and speaker), classify them by presence of noise, correct their transcriptions, and adjust the cutoffs of the signals. Key Considerations:

- **Audio Quality:** Recordings with high SNR (> 20 dB). The audio should be free from background noise and echo, essential to improve the accuracy and naturalness of the TTS. The duration of the clip should be consistent for training, between 3 and 20 s [27].
- **Data Variety:** Samples with different speaking styles, emotional tones, accents, and dialects to improve generalization [4].
- **Amount of Audio:** Between 10 and 20 h of good quality audio. Between 30 and 100 h produce superior results for higher naturalness and precision [12].

Identification, Classification, and Correction. We built a custom web application to select audios for preprocessing; each audio was tagged with a speaker ID, speaker gender, and Quechua variety; audio was classified by noise presence; and we corrected transcriptions and adjusted clip boundaries to ensure higher phrase accuracy. See Fig. 2.

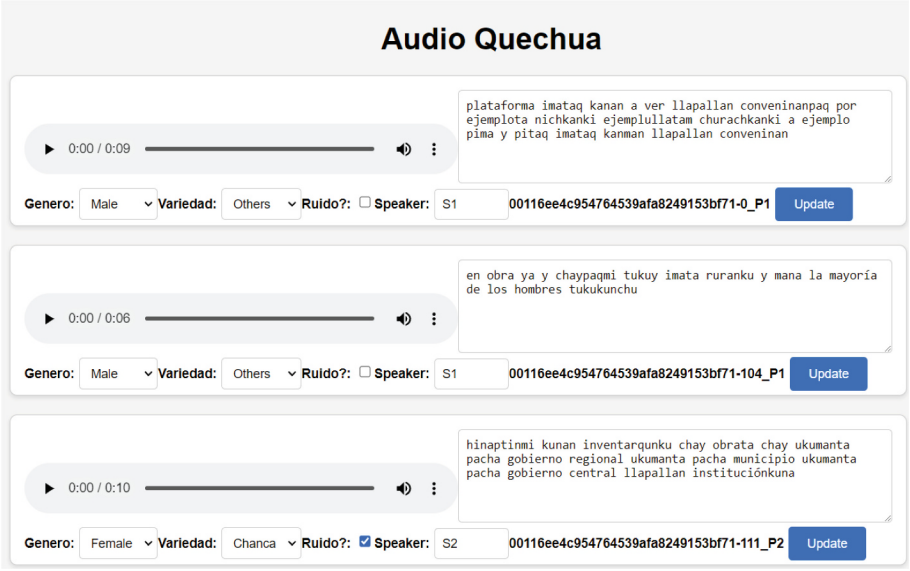


Fig. 2. Web Application: Audio-Quechua.

Audio Quality Enhancement. Because many audios had higher than expected noise, we used MP-SENet [15], an audio enhancement model that processes magnitude and phase spectra in parallel. Based on an embedded codec architecture with a Conformer, it removes noise from both magnitude and phase, significantly improving audio quality. This enhancement was applied to audios tagged as noisy; we improved about 90% of them. Audios that could not be improved were discarded.

2.3 Architectures

TACOTRON 2 + WAVENET [27]. Tacotron 2 is a fully neural TTS system combining a sequence-to-sequence recurrent network with attention to predict mel spectrograms, with a modified WaveNet vocoder. The resulting system synthesizes speech with Tacotron-1-level prosody and WaveNet-level audio quality. It can be trained directly from data without complex feature engineering and achieves state-of-the-art sound quality close to natural speech. However, Tacotron 2 is typically trained with single-speaker data, limiting its ability to synthesize new voices.

FASTPITCH + WAVEGLOW [34]. FastPitch is a fully parallel TTS model that improves on FastSpeech [24] by conditioning the pitch contours, allowing dynamic control of tonal variations in speech. This makes the model more expressive and more aligned with the meaning of the text. It uses two main transformers to process input tokens and output Mel spectrograms, predicting both duration and pitch per segment. Despite being fully parallel, FastPitch reaches the 900× real-time synthesis speed while maintaining quality comparable to autoregressive models.

VITS [12]. VITS combines a conditional variational autoencoder (C-VAE) with an adversarial learning approach, enabling more flexible and natural TTS. This design captures the complex non-linear relationship between textual representations and waveforms, offering the ability to generate multiple speaking styles for the same text input. A stochastic duration predictor enables precise text–audio alignment, while a HiFi-GAN-based discriminator ensures speech quality. VITS enables efficient parallel generation and achieves better subjective quality (MOS) than Tacotron 2 and Glow-TTS.

XTTS v2 [5]. XTTS v2 presents an innovative multilingual TTS model that addresses a critical limitation: functioning effectively in multiple languages. Unlike earlier models typically limited to one or a few languages, XTTS supports 16 languages and can operate zero-shot, imitating new voices with only a few seconds of audio. The architecture, based on Tortoise with significant modifications, combines a 13M-parameter VQ-VAE, a 443M-parameter GPT-2 encoder, and a 26M-parameter HiFi-GAN decoder. Trained on a massive 27,281.6-hour dataset, it achieves state-of-the-art results for most supported languages. A notable feature is its ability to imitate complex speaking styles across languages after training on just one.

3 Experiments

This section describes experiments across four deep learning architectures for Quechua TTS: Tacotron 2 + WaveNet, FastPitch + WaveGlow, VITS and XTTS v2. For each architecture, we monitored key training parameters, including loss function, per-epoch training time, and resource utilization. The results for each architecture are available at https://pucpehuaman.github.io/tts_qu/ (Table 2).

Table 2. Hardware specifications across experiments

Component	Specification
GPU	NVIDIA RTX A5500/4090/A6000
VRAM	24 GB
RAM	64-260 GB
CPU	AMD Threadripper PRO 5955WX/Intel i9-13900K
OS	Ubuntu 24.04.1 LTS/WSL 2.0 Ubuntu 24.0 LTS

3.1 TACOTRON 2 + WAVENET

Tacotron 2 training used 1,453 audios (3.4 h) of female speech. We used NVIDIA's **NeMo** framework [20], leveraging its pre-trained models and TTS optimizations (Table 3).

Table 3. Tacotron 2 + Wavenet configuration

Parameter	Value
Model	Tacotron 2
Sampling rate	22,050 Hz
Mel channels	80
Window size	1024
Window hop	256
Encoder dimension	512
Decoder dimension	1024
Optimization	Adam, init LR 1×10^{-3}
LR schedule	CosineAnnealing, min LR 1×10^{-5}
Batch size	16
Epochs	1000
Training time	48 h

The results did not meet expectations mainly because the model requires more audio from the *same* speaker to better capture vocal characteristics and achieve natural synthesis.

3.2 FASTPITCH + WAVEGLOW

As with Tacotron 2, FastPitch training used 1,453 female audios (3.4 h) and the NVIDIA **NeMo** framework (Table 4).

Table 4. FastPitch + WaveGlow configuration

Parameter	Value
Model	FastPitch
Sampling rate	22,050 Hz
Mel channels	80
Min pitch	65.41 Hz
Max pitch	2093.00 Hz
Symbol embedding dim	384
Optimization	AdamW, init LR 1×10^{-3}
LR schedule	NoamAnnealing
Batch size	32
Epochs	1000
Training time	9.08 h

FastPitch outperformed Tacotron 2 in synthesis quality and significantly reduced the training time to 19% of Tacotron 2's duration.

3.3 VITS

VITS training used 1,453 audios (3.4 h) of female speech. We fine-tuned a Spanish VITS model to obtain a Quechua model using the **Coqui TTS** framework [1] (Table 5).

After 7.7 h, VITS outpaced Tacotron 2 and FastPitch in synthesis quality, with marked improvements in pronunciation clarity, prosodic naturalness, and pitch consistency.

3.4 XTTS V2

For XTTS v2, a ZS-TTS (zero shot text-to-speech) model, we used 937 audios (2.2 h) of female speech. We fine-tuned the Spanish base model using **Coqui TTS** (Table 6).

Figure 3 shows an obvious decrease in eval loss from ~ 0.035 to ~ 0.01, indicating an effective convergence. After 6.9 h, we obtained a model with superior synthesis quality compared to the three previous models.

Table 5. VITS configuration

Parameter	Value
Model	VITS
Sampling rate	22,050 Hz
Mel channels	80
Epochs	2000
Batch size	48
Text cleaner	`multilingual_cleaners`
Use phonemes	`True`
Phoneme language	`es`
Mixed precision	`True`
Training time	7.7 h

Table 6. XTTS v2 configuration

Parameter	Value
Model	XTTS v2.0
Sampling rate	22,050 Hz (in)/24,000 Hz (out)
Learning rate	5×10^{-6}
Optimization	Adam
Batch size	2
Grad Accumulation Steps	84
Audio tokens	1026
Max audio length	∼11.6 s
Max text length	4000 tokens
Training time	6.9 h

4 Results

We present and analyze the results obtained from the experiments described above. Audio and visual materials can be accessed at https://pucpehuaman.github.io/tts_qu/.

4.1 Comparative Analysis

The 28.1M-parameter Tacotron 2 + WaveNet model performed poorly with F0 RMSE of 62 Hz and low perceptual quality (MOS-N: 1.5, MOS-S: 1.4). The 45.7M-parameter FastPitch + WaveGlow showed a moderate improvement with F0 RMSE of 58 Hz and slightly better perceptual scores (MOS-N: 1.9, MOS-S: 1.8). The VITS model with 83M parameters achieved better perceptual quality (MOS-N: 2.1, MOS-S: 2.3) but showed poor pitch accuracy with an F0 RMSE of 120 Hz.

The 518.4M parameter XTTS v2 model performed outstandingly compared to the other three. With an F0 RMSE of 12 Hz, pitch accuracy is significantly superior, indicating advanced capability to replicate natural prosodic patterns. SECS of 0.98 suggests near-perfect stability and highly coherent audio generation. Moreover, MOS-N (4.9) and MOS-S (4.8) show that Quechua-speaking evaluators perceive the speech as nearly indistinguishable from human voice, making it suitable for practical applications (Table 7).

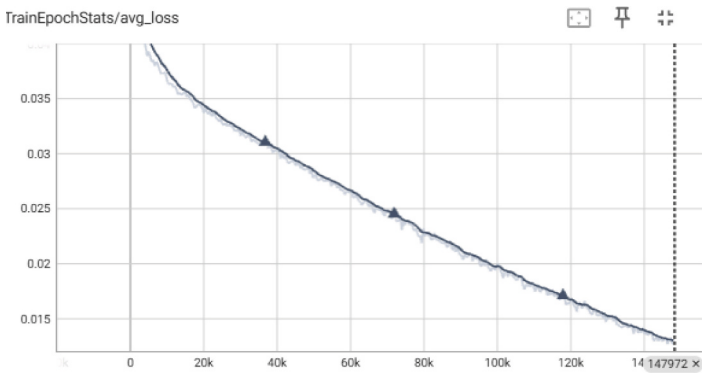


Fig. 3. Evaluation loss—XTTS v2.

Table 7. Comparative Evaluation of Quechua TTS Models

Model	F0 RMSE	SECS	MOS-N	MOS-S
XTTS v2	12 Hz	0.98	4.9	4.8
VITS	120 Hz	0.89	2.1	2.3
FastPitch + WaveGlow	58 Hz	0.88	1.9	1.8
Tacotron 2 + WaveNet	62 Hz	0.88	1.5	1.4

Figure 4 illustrates the comparison F0 between a reference and each model:

- **Tacotron 2 + WaveNet:** The generated F0 shows lower variability than the original, failing to capture the original variations.
- **FastPitch + WaveGlow:** Slight improvement with more dynamism, though there are notable differences in precision and variability in some segments.
- **VITS:** Significant discrepancy, with consistently higher F0 than the original, reflecting considerable inaccuracy.
- **XTTS v2:** Clearly superior, faithfully replicating fluctuations and overall contours of the original audio; best in precision and similarity.

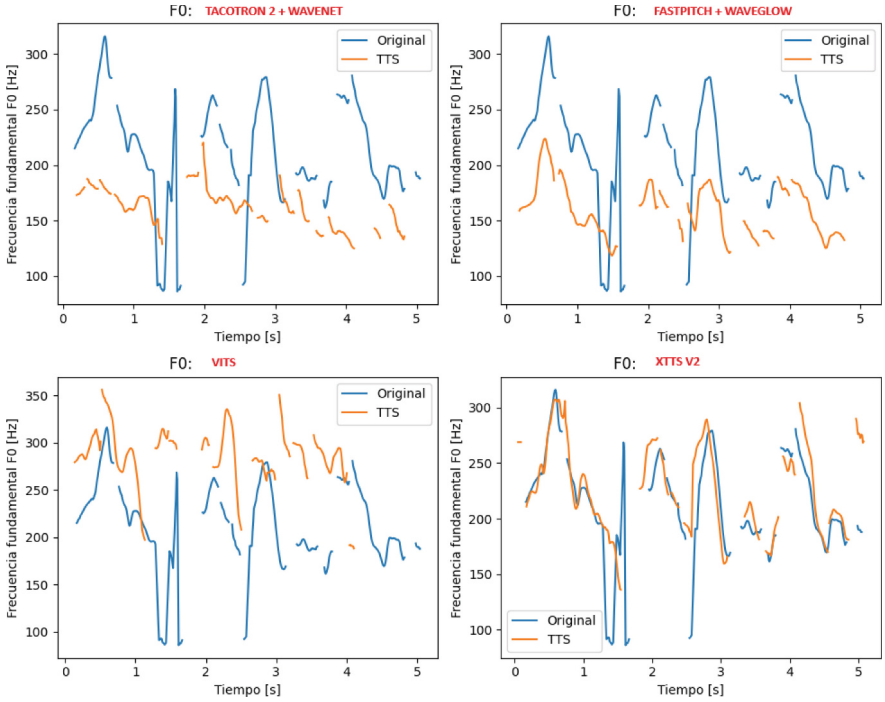


Fig. 4. Visual comparison of F0 for each trained model.

5 Conclusions and Future Work

5.1 Conclusions

The end-to-end XTTS v2 model clearly outperforms both two-stage models (Tacotron 2 + WaveNet and FastPitch + WaveGlow) and the end-to-end VITS, excelling at replicating natural prosodic patterns, generating highly coherent audio, and achieving superior perceived naturalness and similarity. Quantitatively, XTTS v2 achieved the best performance in objective metrics (F0 RMSE and SEC). The subjective metrics (MOS-N and MOS-S) also confirmed that end-to-end models provide more natural, fluent synthesis than two-stage ones.

Data preprocessing was crucial for successful training, allowing a clearer and more uniform representation of speech features, directly reflected in the quality and consistency of experimental results. The evaluation of style and prosody showed that XTTS v2 captures subtler pitch and intonation variations, surpassing the other models. The study also demonstrates that the quality of the data set is as important as quantity; careful curation and cleaning were the key to highly satisfactory results, especially with end-to-end models such as VITS and XTTS v2. Moreover, transfer learning through fine-tuning from Spanish models is efficient and effective for low-resource languages, offering methodological guidance for future research in languages similar to Quechua.

Finally, this research highlights AI's potential to reduce digital divides and promote linguistic and technological inclusion. The quality achieved by XTTS v2 allows immediate integration into everyday technological platforms, facilitating access to information in Quechua and promoting both prestige and active use of the language.

The fine-tuned XTTS v2 model (Quechua–Chanka) and training configurations are publicly available at the Hugging Face repository: huggingface.co/edhuc/qcr_xttsv2. These artifacts are designed for teachers, community media, and ASR teams, lowering operational barriers to producing Quechua audio at scale.

5.2 Future Work

We propose expanding the data set with recordings covering different Quechua dialects, male and female voices, diverse ages, and speaking styles (narration, informal conversation, emotional expressions) to improve generalization and robustness when training XTTS v2 or other architectures. In parallel, we will apply data enhancement techniques to expand training and evaluation sets [22]. We will also optimize XTTS v2 for low-resource devices, especially mobile. Finally, integrating XTTS v2 with advanced Quechua ASR will enable full voice interfaces capable of receiving spoken instructions and responding in real time in Quechua, powering virtual assistants and multilingual conversational applications accessible to Quechua-speaking communities.

References

1. AI, C.: Coqui tts (2024). https://github.com/coqui-ai/TTS, repositorio GitHub, consultado el 12 de marzo de 2025
2. Alzate López, H.: Física de las ondas. Editorial Universidad de Antioquia, Medellín, Colombia (2007). https://bibliotecadigital.udea.edu.co/bitstream/10495/1550/13/AlzateLopezHector_2007_FisicaOndas.pdf
3. Cardenas, R., Zevallos, R., Baquerizo, R., Camacho, L.: Siminchik: a speech corpus for preservation of Southern Quechua. In: Proceedings of the 2nd Workshop on Improving Social Inclusion using NLP: Tools, Methods, and Resources (ISINLP2), co-located with LREC 2018, pp. 21–26. European Language Resources Association (ELRA), Miyazaki, Japan (2018). http://lrec-conf.org/workshops/lrec2018/W14/pdf/4_W14.pdf
4. Casanova, E., et al.: YourTTS: towards zero-shot multi-speaker TTS (2022)
5. Casanova, E., et al.: XTTS: a massively multilingual zero-shot text-to-speech model (2024)
6. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Springer (1997)
7. Facultad de Letras y Ciencias Humanas: La inteligencia artificial ya habla en idioma quechua (2023). https://letras.unmsm.edu.pe/noticias/la-inteligencia-artificial-ya-habla-en-idioma-quechua/
8. González-Zepeda, L., Martínez-Pinto, C.: Inteligencia artificial centrada en los pueblos indígenas: Perspectivas desde américa latina y el caribe. Tech. rep., UNESCO, París y Montevideo (2023). http://creativecommons.org/licenses/by-sa/3.0/igo/

9. Hornberger, N.H., Coronel-Molina, S.M.: Quechua language shift, maintenance, and revitalization in the Andes: the case for language planning (2004). https://doi.org/10.1515/ijsl.2004.025
10. International Telecommunication Union: Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union, Geneva, Switzerland (August 1996), iTU-T Recommendation P.800
11. Johnson, K.: Acoustic and Auditory Phonetics, 3 edn. Wiley-Blackwell (2012)
12. Kim, J., et al.: VITS: conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: ICML (2021)
13. Li, Y.A., Han, C., Raghavan, V.S., Mischler, G., Mesgarani, N.: StyleTTS 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models (2024)
14. Liu, E.M., Yeh, J.W., Lu, J.H., Liu, Y.W.: Speaker embedding space cosine similarity comparisons of singing voice conversion models and voice morphing. In: Proceedings of the Journal of the Acoustical Society of America, vol. 154, p. A244. Acoustical Society of America (2023). https://doi.org/10.1121/10.0023424, https://pubs.aip.org/asa/jasa/article/154/4_supplement/A244/2924394/Speaker-embedding-space-cosine-similarity
15. Lu, Y.X., Yang, A.F., Ling, Z.H.: MP-SENet: a speech enhancement model with parallel denoising of magnitude and phase spectra (2023)
16. Meyer, J.: Speech technology for endangered languages. Computer Speech & Language (2022)
17. Ministerio de Cultura del Perú: Política Nacional de Lenguas Originarias, Tradición Oral e Interculturalidad al 2040. Ministerio de Cultura del Perú (2017). https://repositorio.cultura.gob.pe/handle/CULTURA/6715, aprobada mediante el Decreto Supremo N.o 005-2017-MC
18. Moore, B.: An Introduction to the Psychology of Hearing. Brill (2013)
19. Moseley, C. (ed.): Atlas of the World's Languages in Danger. UNESCO Publishing, Paris, 3 edn. (2010). https://unesdoc.unesco.org/ark:/48223/pf0000189453
20. NVIDIA Corporation: NVIDIA NeMo Framework: Text-to-Speech User Guide. NVIDIA (September 2024). https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/tts/intro.html, accedido el 11 de marzo de 2025
21. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing. Prentice Hall (1999)
22. Park, D., et al.: SpecAugment: a simple data augmentation method. In: Interspeech (2019)
23. Proakis, J.G., Manolakis, D.G.: Digital Signal Processing. Pearson (2006)
24. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. arXiv preprint arXiv:1905.09263 (2019)
25. Renker, T.A.: El uso del quechua en el Perú: Una investigación de identidad y performance cultural (2014)
26. Romero, M., Gómez-Canaval, S., Torre, I.: Automatic speech recognition advancements for indigenous languages of the Americas. Appl. Sci. **14**, 6497 (2023)
27. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions (2018)
28. Stevens, S.S., Volkmann, J., Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. **8**(3), 185–190 (1937)
29. Tan, X., et al.: A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561 (2021)

30. Vargas Muñoz, J.E., Cruz Tello, J.A., Castro Mamani, R.A.: Let's speak Quechua: the implementation of a text-to-speech system for the incas' language. In: IberSPEECH 2012 – VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, pp. 300–307. Universidad Nacional de San Antonio Abad del Cusco, Madrid, España (2012)
31. Wang, Y., et al.: Tacotron: towards end-to-end speech synthesis (2017)
32. Zevallos, R., Camacho, L., Melgarejo, N.: Data augmentation for low-resource Quechua ASR improvement. arXiv preprint (2022). https://doi.org/10.48550/arXiv.2207.06872, https://arxiv.org/abs/2207.06872
33. Zevallos, R., Camacho, L., Melgarejo, N.: Huqariq: a multilingual speech corpus of native languages of Peru for speech recognition. arXiv preprint arXiv:2207.05498 (2022), https://arxiv.org/abs/2207.05498
34. Łańcucki, A.: FastPitch: parallel text-to-speech with pitch prediction, Tech. rep. NVIDIA Corporation (2021)

Quechua Speech Datasets in Common Voice: The Case of Puno Quechua

Elwin Huaman[1(✉)], Wendi Huaman[2], Jorge Luis Huaman[2], and Ninfa Quispe[2]

[1] University of Innsbruck, Innsbruck, Austria
elwin.huaman@uibk.ac.at
[2] National University of Altiplano Puno, Puno, Peru

Abstract. Under-resourced languages, such as Quechuas, face data and resource scarcity, hindering their development in speech technology. To address this issue, Common Voice presents a crucial opportunity to foster an open and community-driven speech dataset creation. This paper examines the integration of Quechua languages into Common Voice. We detail the current 17 Quechua languages, presenting Puno Quechua (ISO 639-3: qxp) as a focused case study that includes language onboarding and corpus collection of both reading and spontaneous speech data. Our results demonstrate that Common Voice now hosts 191.1 h of Quechua speech (86% validated), with Puno Quechua contributing 12 h (77% validated), highlighting the Common Voice's potential. We further propose a research agenda addressing technical challenges, alongside ethical considerations for community engagement and indigenous data sovereignty. Our work contributes towards inclusive voice technology and digital empowerment of under-resourced language communities.

Keywords: Quechua Languages · Common Voice · Speech Datasets

1 Introduction

The digital landscape for indigenous languages, especially those with limited written traditions, presents challenges and opportunities. A well-know challenge in language technology is the lack of data and linguistics resources in under-resourced languages [3,5], such as the Quechua languages [6]. At the same time, initiatives such as Mozilla's Common Voice project offer an opportunity to create and expand speech datasets for these languages [2].

Among the +40 Quechua language varieties identified by ISO 693-3[1] and Ethnologue[2] as Active, Stable, or Institutional, only four are taken into account by the Government of Peru [9]. Even these are not treated as distinct ISO codes, but rather as regional groupings of multiple varieties [9]. Linguistically, there

[1] https://iso639-3.sil.org/code/que.
[2] https://www.ethnologue.com/language/que/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 184–193, 2026.
https://doi.org/10.1007/978-3-032-20322-9_13

are two major groups of Quechuas: Quechua I and Quechua II, and several subgroups [7,10], and most of these Quechua-speaking populations are located within Peruvian territory. According to the 2017 National Census in Peru [9], the country has approximately 3805531 Quechua speakers, and in the Puno region alone there are 474203 Quechua speakers.

This paper aims to present an overview of Quechua language speech datasets within the Mozilla Common Voice platform[3] and to detail the integration of the Puno Quechua (qxp) into the platform. Furthermore, we explore its current statistical landscape, the Puno Quechua integration, and discuss a research agenda related to ethical considerations and community engagement.

The remainder of the paper is organized as follows. Section 2 provides an overview of Quechua languages in the Common Voice platform. Section 3 and 4 describe the process and results of integrating Puno Quechua. Finally, Sect. 5 concludes the paper with some remarks and future work plans.

2 Quechua Languages in Common Voice

This section discusses the integration of Quechua languages into Common Voice, highlighting the platform's collaborative framework, Quechua linguistic diversity, and current status of contributions (see Fig. 1).

2.1 Common Voice: A Collaborative Framework for Speech Data

The Common Voice project is an initiative to make speech technology accessible to everyone, especially to under-resourced languages. It started in 2017 with an emphasis on the English language, since 2018 was made available to any language. The operational framework is entirely reliant on volunteer contributions: users collect and upload sentences, then sentences are recorded, afterwards sentences are validated by others. This process allows anyone to contribute with more sentences, recordings, and validations [2], ensuring a quality control over the data. The resulting dataset consists of MP3 files with their corresponding texts, if available, including demographic metadata like age, sex, and accent.

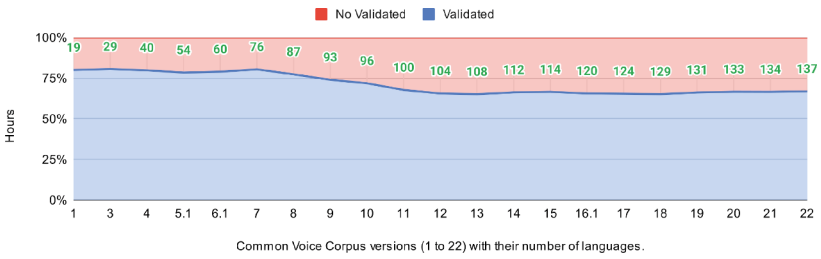


Fig. 1. Language coverage by Common Voice versions.

[3] https://commonvoice.mozilla.org/.

Since its inception, Common Voice has grown significantly. The Common Voice corpus expanded from 1368 h across 19 languages in version 1 to 33815 h across 137 languages in version 22[4], see Fig. 1. Starting from version 17, sentences are categorized into 11 thematic domains, see Fig. 2. All collected data is released under a Creative Commons CC0 license, which fosters development and innovation in voice-enabled technologies.

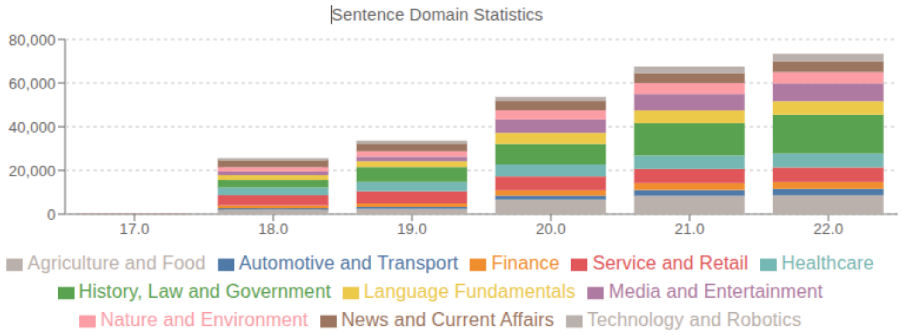


Fig. 2. Sentence domain distribution in Common Voice versions.

Integrating a Language Into Common Voice. It involves two main steps:

1. **Language Onboarding**: A language is proposed via GitHub by opening an issue with language-specific details[5]. Once it is approved, a localization project is created on Pontoon[6] to translate the Common Voice platform.
2. **Corpus Collection**: A set of sentences with CC0 licesing is required for people to read aloud. After this is uploaded and approved, the language is officially launched on the Common Voice platform.

Once integrated, community members can begin recording and validating voice clips. Language data is typically released as part of the Common Voice corpus within three months[7].

2.2 Quechua Languages

Quechua languages, 43 identified on Ethnologue[8] and 44 identified on Glottolog[9], are deeply rooted in South America. Quechua speakers are distributed

[4] https://github.com/common-voice/cv-dataset/blob/main/README.md.
[5] https://github.com/common-voice/common-voice.
[6] https://pontoon.mozilla.org/.
[7] https://commonvoice.mozilla.org/en/about.
[8] https://www.ethnologue.com/language/que/.
[9] https://glottolog.org/resource/languoid/id/quec1387.

across Argentina, Bolivia, Chile, Colombia, Ecuador, and Peru, with an estimated 7 million native speakers overall, and in Peru alone 3805531 quechua speakers [9]. Currently, 17 Quechua languages are supported on Common Voice, each identified by an ISO 639-3 Code[10]. These languages are at various stages of integration: language onboarding or corpus collection, see Sect. 2.1.

Figure 3 shows localization efforts by Quechua-speaking communities to integrate their language into Common Voice. Most languages have localized the minimum 20% localization threshold required for basic functionality and reading-based speech recording. However, 75% is the minimum required to make the user-platform interaction more correct and to support spontaneous speech recording.

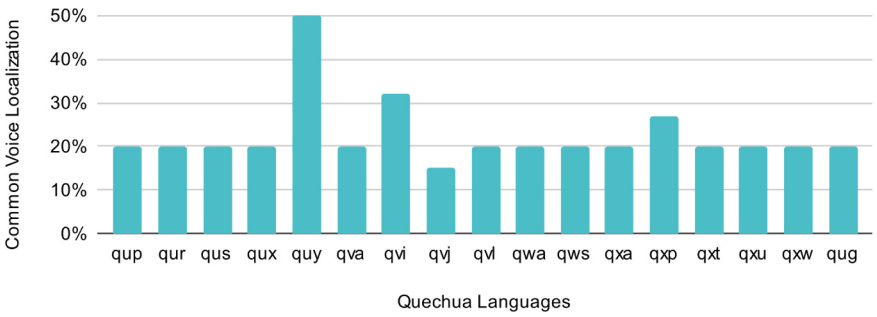


Fig. 3. Localization efforts by Quechua communities in Common Voice platform.

The quantity of recorded speech varies widely among Quechua languages. For example, qug (Chimborazo Highland Quichua) has 0 h recorded, while qup (Southern Pastaza Quechua) has 16 h, see Fig. 4. Overall, approximately 86% of Quechua submitted recordings have been validated.

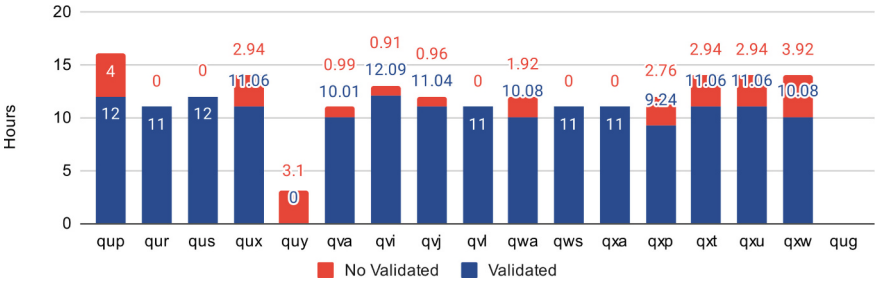


Fig. 4. Speech hours by Quechua language in Common Voice.

[10] https://iso639-3.sil.org/code/que.

3 Case Study: Puno Quechua

This section presents a case study on the integration of Puno Quechua into the Common Voice platform. It begins by characterizing Puno Quechua linguistically and demographically, followed by a detailed account of the steps taken to integrate it into Common Voice, including platform localization, sentence collection, and speech recording efforts.

3.1 Puno Quechua Language

Puno Quechua, identified by the ISO 639-3: qxp[11], belongs to the Quechua II group. According to the Peruvian government, it is further sub-grouped under Southern Quechua, specifically within the Quechua Collao. The 2017 National Census [9] estimates approximately 474203 speakers of Puno Quechua, predominantly in the Puno region of Peru. Historically, the number of Quechua speakers in this region has shown a significant decline: from 87.23% in 1940 to 50% in 1993, and further to 38.01% in 2007 [1,4,8].

According to Glottolog, most linguistic resources for Puno Quechua were published between 1963 and 1993, with few notable exceptions from 2005. These resources offer valuable insights into the language's linguistic characteristics. Linguistically, Puno Quechua adheres to the agglutinative structure characteristic of all Quechua languages, where words are formed by adding multiple suffixes to a root word, and follows a Subject-Object-Verb (SOV) sentence structure.

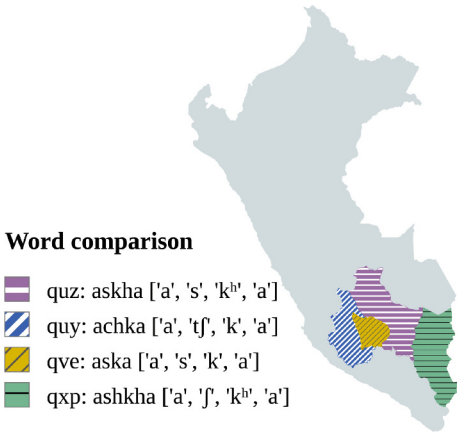


Fig. 5. Comparison of a word across Southern Quechua languages.

A comparative lexical analysis across different Southern Quechua varieties, as illustrated in Fig. 5, demonstrates lexical variation[12]. For instance, the English

[11] https://iso639-3.sil.org/code/qxp.

[12] Cross-Linguistic Data Formats dataset https://cldf.clld.org/.

word *a lot* is represented as *ashkha* in Puno Quechua (qxp), but varies in other Southern Quechuas. This lexical diversity underscores the need for separate speech datasets for each Quechua language, ensuring linguistic accuracy and utility for speech technology development.

3.2 Integrating Puno Quechua into Common Voice

The integration of Puno Quechua into Common Voice involved two tasks: language onboarding and corpus collection.

Language Onboarding: This task entailed requesting our language via GitHub, where we opened an issue with Puno Quechua language information. Once it was approved, we got access to a Common Voice Pontoon project, where we started translating and localizing the interface into Puno Quechua. Our efforts resulted in a 27% translation rate, which indicates that 466 out of 1793 total strings have been translated[13]. This partial localization covers both Reading Speech and Spontaneous Speech interfaces.

Corpus Collection: This phase focused on gathering CC0-licensed sentences suitable for the platform. We collected 2065 sentences and subsequently uploaded them to the Common Voice platform. These Puno Quechua sentences represent 11.6% of the total Quechua sentences available on the platform (see Fig. 6).

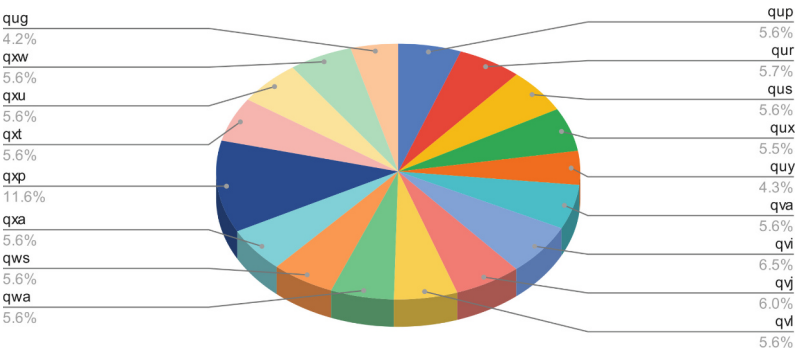


Fig. 6. Quechua sentences coverage by Quechua languages.

3.3 Speech Recording

Speech recording efforts for Puno Quechua within Common Voice were divided into Reading Speech and Spontaneous Speech contributions.

[13] https://pontoon.mozilla.org/qxp/common-voice/.

Reading Speech Recording: It involved volunteers reading aloud the uploaded sentences and recording them. We try our best to contribute with clear pronunciation and in noise-free environment, so the recordings are of high quality. So far, community validation has reached 77% (9 h and 24 min) of the total recorded clips (12 h), see Fig. 4. We highlight the importance of sustained community engagement to improve the quality and robustness of the Puno Quechua speech dataset.

Spontaneous Speech Recording: This more advanced contribution type involves community members answering open-ended questions, which are then transcribed and validated. We uploaded 40 questions specifically within the Agricultural and Food domain for the community to answer. The transcription process for spontaneous speech also accounts for casual speaking patterns, including interjections like "mm" and "ahh", as well as repetitions. Subsequently, these transcriptions are validated against the speech recordings by other speakers. This approach captures diverse speaking styles, including accents, hesitations, and informal phrasing, which are vital for developing robust and adaptable voice technologies.

4 Results

This chapter presents the key outcomes of integrating Quechua languages, with a specific focus on Puno Quechua, into the Common Voice platform. It highlights the current status of the collected speech corpus and outlines a comprehensive research agenda and social framework necessary for advancing the development of high-quality speech datasets for under-resourced languages.

4.1 Current Status of Quechua Languages in Common Voice

Following the integration of 17 Quechua languages into the Common Voice platform, particularly the Puno Quechua, this paper demonstrates a viable pathway for making under-resourced languages available as community-driven text-to-speech corpora. Currently, the Common Voice platform hosts 191.1 h of speech data across all integrated Quechua languages, of which 163.72 h (86%) have been validated by the community. In the case of Puno Quechua, a total of 12 h of speech were recorded, with 77% of this data successfully validated. Table 1 provides a summary of the current contributions.

Table 1. Overview of Quechua Languages Data on Common Voice.

	Hours	Speakers	Validation Progress	Sentences
Quechua Languages	191.1	238	86%	17754
Puno Quechua	12.0	14	77%	2065

4.2 Research Agenda for High-Quality Quechua Speech Datasets

To ensure the creation of high-quality Quechua speech datasets suitable for diverse applications such as speech recognition, text-to-speech (TTS) synthesis, and language preservation, research challenges and opportunities must be addressed. This research agenda outlines, integrating the technical requirements of the Common Voice platform, the linguistic complexities of under-resourced languages, and the socio-cultural dynamics required for community engagement.

Challenge: Language Support and Orthographic Standards. It involves ensuring that all languages, regardless of their listing on Glottolog or ISO, are adequately supported. This requires review to ensure localizations and collected sentences adhere to established orthographic standards. **Opportunity**: Collaborating with universities, research institutes, and Quechua-speaking organizations can facilitate the quality integration of any language into Common Voice by leveraging their linguistic expertise and community ties.

Challenge: Building Diverse and Culturally Relevant Text Corpora. This includes ensuring a balanced distribution of sentences across various domains to mitigate bias in the resulting speech dataset. **Opportunity**: Actively mobilizing the community in the sentence collection process is crucial for ensuring that gathered sentences are culturally appropriate and domain-relevant, thereby reducing inherent biases.

Challenge: High-Quality Voice Contributions. It involves effective training and continuos support for community contributors. Practical barriers identified during Puno Quechua data collection include expensive internet access and limited digital literacy, which hinder community participation. **Opportunity**: A hybrid approach combining online Common Voice contributions with organized offline recordings campaigns is essential. This strategy should include the provision of devices, internet access, offline access to the platform, alongside localized training workshops, to maximize participation and overcome socio-economic barriers.

Challenge: Spontaneous Speech and Code-Switching. It involves capturing spontaneous speech "on the fly", which brings unique challenges due to linguistic phenomena that deviate from carefully read text, such as prolonged pauses, false starts, repetitions, and background noise. An additional complexity arises from the frequent code-switching between Quechua and Spanish among bilingual speakers. **Opportunity**: The development of offline mobile applications can democratize data collection, enabling participation from a wider range of speakers and in diverse contexts. This directly addresses the need for diverse datasets to reduce bias in speech recognition models. For Puno Quechua, we focus our effort to minimize code-switching by collecting speech centered around culturally vibrant topics such as agriculture, farming, and daily life.

4.3 Social Framework and Ethical Considerations

Beyond the technical research agenda, the sustainable integration of under-resourced languages into platforms like Common Voice necessitates a robust social framework grounded in ethical considerations:

- **Prioritize Localization and Community Building:** Initial localization efforts for activating a language on the platform. Simultaneously, it is necessary to build relationships with Quechua communities, linguists, research institutions, and public organizations to foster sustainable engagement.
- **Ethical Framework and Data Sovereignty:** It is essential to establish a solid ethical framework that goes beyond mere regulatory compliance. This includes implementing informed consent processes that are culturally sensitive. Furthermore, it is crucial to engage with language communities to ensure a comprehensive understanding of the Creative Common CC0 licensing towards their data sovereignty. This engagement should explore the potential for community-centric data licenses or benefit sharing agreements that align with community values.
- **Importance of Spontaneous Speech:** The inclusion of spontaneous speech is crucial for training real-world speech recognition applications, as most voice assistants require the ability to understand natural and unscripted speech to be effective in daily use.

By addressing the outlined research agenda and ethical considerations, the integration of languages into Common Voice can establish a robust foundation, empowering language communities to contribute significantly. This paper brings insights into the digital preservation and revitalization of under-resourced languages through their integration into the Common Voice ecosystem.

5 Conclusions

This paper demonstrates the integration of under-resourced Quechua languages into Mozilla's Common Voice platform, showcasing a viable model for creating vital speech datasets. We highlighted the contribution of 191.1 h of Quechua speech (86% validated), including 12 h for Puno Quechua (77% validated), underscoring Common Voice's potential to address speech data scarcity.

To advance these efforts, a comprehensive research agenda is crucial, focusing on ensuring linguistic accuracy through distinct datasets, developing diverse text corpora, and overcoming digital access barriers via hybrid online-offline approaches for high-quality, spontaneous speech collection. Furthermore, a robust social framework is essential, emphasizing community building, culturally sensitive ethical practices, and discussions on indigenous data sovereignty.

Our future work will be focused on expanding the Common Voice dataset to include all remaining Quechua languages, while ensuring quality, diversity, and sustainability throughout community contributions. Additionally, we will explore ways to enable offline contributions through integrations with instant messaging applications, allowing people to contribute from different devices and connection modes.

Acknowledgments. We thank the Mozilla Foundation, the Common Voice project, and the Mozilla Common Voice community for their support. A deep gratitude to Francis Tyers, for his unconditional support in overcoming challenges during the Speech recording campaigns. Our sincere appreciation to the T'ikariy community in Nuñoa, Puno, Peru, specially Christian Tapara, Eliana Quispe, Luis Huaman, Magaly Mullisaca, Maribel Hancco, Milagros Huaman, Nely Quispe, and everybody who contributed with their voices to this project.

References

1. Andrade Ciudad, L.: Ten news about quechua in the last peruvian census. Letras (Lima) **90**, 41–70 (2019). https://doi.org/10.30920/letras.90.132.2
2. Ardila, R., et al.: Common voice: a massively-multilingual speech corpus. In: Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, pp. 4218–4222. European Language Resources Association (2020). https://aclanthology.org/2020.lrec-1.520/
3. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Commun. **56**, 85–100 (2014). https://doi.org/10.1016/j.specom.2013.07.008
4. Carbajal Solis, V., et al.: Lenguas Originarias del Peru. Ministerio de Educacion del Peru (2019). https://hdl.handle.net/20.500.12799/6261
5. Huaman, E., Huaman, J.L., Huaman, W.: Getting quechua closer to final users through knowledge graphs. In: Information Management and Big Data - 9th Annual International Conference, SIMBig 2022, Lima, Peru, November 16-18, 2022, Proceedings. Communications in Computer and Information Science, vol. 1837, pp. 61–69. Springer (2022). https://doi.org/10.1007/978-3-031-35445-8_5
6. Huaman, E., Huaman, W., Huaman, J.L.: Making an under-resourced language available on the wikidata knowledge graph: quechua language. In: Information Management and Big Data - 9th Annual International Conference, SIMBig 2024, Ilo, Peru, November 20-22, 2024, Proceedings. Communications in Computer and Information Science, vol. 2496, pp. 212–224. Springer (2024). https://doi.org/10.1007/978-3-031-91428-7_15, https://doi.org/10.1007/978-3-031-91428-7_15
7. Huaman, E., Lindemann, D., Caruso, V., Huaman, J.L.: QICHWABASE: a quechua language and knowledge base for quechua communities. CoRR **abs/2305.06173** (2023). https://doi.org/10.48550/ARXIV.2305.06173
8. Lópes, L.E.: Educación bilingüe en puno-perú: activos y pasivos de un programa de educatión rural en los andes. Lingüística indígena e educação na América Latina. Campinas: UNICAMP, pp. 13–70 (1993, https://etnolinguistica.wdfiles.com/local--files/biblio%3Alopez-1993-educacion/Lopez_1993_EducacionBilinguePunoPeru.pdf
9. Ministerio de Educación del Perú: Lenguas originarias del Perú. Biblioteca Nactional del Perú, Lima, Perú (2018)
10. Torero, A.: Los dialectos quechuas. In: Anales Cientificos, Volume II, Lima, 1964, pp. 446–478. Universidad Agraria Peru (1964)

Detecting Explicit Content in Spotify Song Lyrics Using Natural Language Processing Techniques

Aaron C. A. Navarro Mendoza, Dylan J. Huarcaya Farfan(✉), Fernando A. Usurin Arias, Marcelino Maita-Vargas, and Ariana M. Villegas Suarez

University of Engineering and Technology - UTEC, Barranco 15063, Peru
{aaron.navarro,dylan.huarcaya,fernando.usurin,
marcelino.maita,ariana.villegas}@utec.edu.pe

Abstract. The proliferation of music streaming platforms has created an urgent need for automatic explicit content detection to protect vulnerable audiences, particularly children, from inappropriate material. This study presents a comprehensive comparative analysis of machine learning approaches for classifying explicit content in song lyrics using a dataset of 108,138 songs from Spotify. We evaluate traditional machine learning models (Decision Tree, Logistic Regression, Naive Bayes, SVM) with two distinct text representation methods: TF-IDF vectorization and pre-trained LyricsBERT embeddings. Additionally, we assess the impact of incorporating musical metadata and acoustic features alongside textual data. Our experimental framework compares feature combinations ranging from lyrics-only to comprehensive multi-modal representations, including one that excludes lyrical data entirely. Results demonstrate that traditional models with TF-IDF vectorization achieve competitive performance, with Decision Tree achieving the highest F1-score of 0.7896 when combining textual, numerical, and categorical features. Surprisingly, metadata-only approaches using Logistic Regression achieved superior performance (F1-score: 0.8715), suggesting that acoustic and categorical features may be more reliable indicators of explicit content than lyrics alone. Deep learning approaches using BERT fine-tuning achieved lower performance (F1-score: 0.7046) at a significantly higher computational cost. These findings challenge the assumption that lyrical content is the primary determinant of explicitness, providing practical insights for developing scalable content moderation systems on streaming platforms.

Keywords: Explicit content detection · Music classification · Natural language processing · TF-IDF · BERT · Content moderation

A. C. A. Navarro Mendoza, D. J. Huarcaya Farfan, F. A. Usurin Arias and M. Maita-Vargas—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 194–206, 2026.
https://doi.org/10.1007/978-3-032-20322-9_14

1 Introduction

Every day, millions of new songs from around the world are added to music platforms like Spotify, YouTube Music, and Tidal. Many of these songs contain explicit lyrics, including sexual references, coarse language, violence, or substance use. However, this screening process is not always accurate, allowing some content to appear in playlists without proper identification, even though these platforms often use an "explicit" warning label.

Vulnerable audiences, such as children, can be involuntarily exposed to this type of content without proper filters, which can have a detrimental effect on their behavior. While warning systems exist, such as the Parental Advisory Label (PAL), created by the Recording Industry Association of America (RIAA), their use is restricted to the catalogs of affiliated artists, limiting their scope [1]. Independent musicians can upload songs to authorized distributors who may ignore this requirement.

The main objective of our project is to address this challenge by automatically analyzing a song's lyrics and other relevant textual elements to determine whether they are explicit or not. This approach aligns with current research focused on developing more sophisticated and accurate methods for detecting explicit content. These systems provide an additional layer of contextual understanding and adaptive audience protection. These systems, for example, can not only determine whether a song is explicit but also categorize the root cause of that explicitness (such as strong language, sexual references, substance abuse, violence, or discrimination) using multi-class labels.

Furthermore, research has shown that simple methods, such as average word embeddings, can outperform more complex models in content labeling tasks, such as explicit content identification. This suggests that an efficient textual representation can be more scalable and effective than deep neural network architectures.

The remainder of this paper is organized as follows. Section 2, 'Related Work', reviews prior studies on explicit content detection in song lyrics. Section 3, 'The Study', describes the dataset used in this research and outlines the exploratory analysis conducted. Section 4, 'Experimentation and Results', presents the models evaluated and their corresponding outcomes. Section 5, 'Discussion', provides an interpretation of the findings and their implications. Finally, Sect. 6, 'Conclusions', summarizes the main contributions of the study and highlights directions for future research.

2 Related Work

Research on the automatic detection of explicit content in song lyrics has advanced notably in the last decade, driven by the expansion of digital music platforms such as Spotify, YouTube Music, and Tidal, alongside progress in natural language processing (NLP). Despite these advances, several challenges remain, including those inherent to the nuances of musical language: cross-cultural and multilingual variation, the frequent use of metaphor and

euphemism, subjective definitions of 'explicitness', and severe class imbalance, as explicit tracks often represent a small fraction of catalogs [2] [3]. These factors can compromise model accuracy and limit its generalizability in real-world applications.

Early approaches relied on lexical methods such as Bag of Words and Term Frequency–Inverse Document Frequency (TF–IDF) combined with classifiers like Support Vector Machines or Random Forests. Such methods, though simple, have shown competitive performance in controlled datasets, especially when dealing with noisy labels such as platform-generated 'Explicit' tags [4,5]. However, their inability to reliably detect indirect or metaphorical references limits their scope, particularly in genres rich in figurative language, such as rap or urban music [6].

The adoption of neural architectures, particularly transformers like BERT and its language-specific variants, has improved the ability to capture contextual meaning and long-range dependencies, often outperforming traditional models when applied to well-curated and balanced corpora [7]. Yet, evidence shows that model sophistication alone does not guarantee higher performance if data quality issues such as inconsistent labeling or imbalance persist [3].

Hybrid and ensemble strategies that combine lexical features with contextual embeddings and diverse classifiers have recently emerged as a robust alternative, improving stability and F1-score in multilingual contexts [8]. Nevertheless, their success depends heavily on appropriate class balancing techniques and careful calibration of decision thresholds.

More recently, the literature has started to discuss the role of Large Language Models (LLMs), which extend transformer architectures to billions of parameters and are capable of zero-shot or prompt-based classification. Rospocher and Eksir [7] explicitly suggest future research directions should compare state-of-the-art multi-label classifiers and investigate prompt-based methods (e.g., InstructGPT, ChatGPT) for explicit content detection. While these approaches have not yet been systematically applied to lyrics, related work in other areas, such as hate speech detection, demonstrates that LLMs can capture subtle figurative and implicit expressions that traditional models often miss [11]. This suggests that the results reported, particularly the superiority of metadata features over lyrics in some studies, might shift significantly when evaluated with contemporary LLM-based methods.

Finally, the literature highlights shortcomings in evaluation practices. Random train-test splits risk inflating performance scores due to data leakage, particularly when songs from the same artist appear in both sets. More reliable protocols include out-of-genre, out-of-time, or artist-stratified evaluations [4,12]. Moreover, while most research adopts a binary classification, applications in parental control or recommendation systems could benefit from multi-tiered taxonomies that distinguish between different forms and severities of explicitness.

In summary, no single approach dominates across all contexts: traditional TF-IDF-based models remain valuable when data are scarce or noisy. At the same time, transformer-based architectures excel in high-quality, large-scale datasets.

The integration of these methods, combined with rigorous evaluation, nuanced classification schemes, and the exploration of LLM-based approaches, appears to offer the most promising path forward.

3 The Study

3.1 Dataset Structure

The dataset employed in this research is `spotify_dataset.csv`, publicly available on the Kaggle platform. It contains information on over 500,000 songs, including acoustic attributes, metadata, lyrics, popularity scores, and an `explicit` flag. For this study, a preprocessing phase was conducted to ensure data quality and computational efficiency.

The main cleaning steps were:

- **Popularity filtering:** Songs with popularity scores below 40% were removed due to their limited relevance and potential incomplete data.
- **Artist unification:** Songs featuring multiple artists were excluded to simplify the analysis and avoid ambiguities.
- **Duplicate correction:** Song and artist names were normalized (lowercase, no punctuation) to identify and remove duplicates (e.g., "It's My Life" vs "Its my life").

After cleaning, the dataset comprised 108,138 songs and 41 attributes, including numerical variables (`tempo`, `loudness`, `energy`, `danceability`), categorical variables (`genre`, `explicit`), and textual variables (song lyrics in `text`).

3.2 Dataset Relevance and Exploratory Analysis

The exploratory analysis revealed that 23.9% of the songs are explicit, while 76.1% are non-explicit (Fig. 1). This indicates a predominance of general-audience content, although explicit songs still represent a significant proportion.

Correlation analysis (Fig. 2) reveals that among all features, `danceability` exhibits the strongest positive linear correlation with explicit content, followed by `loudness (dB)`. This suggests that explicit songs are not only more likely to be danceable but also tend to be louder compared to non-explicit tracks.

To further investigate beyond linear dependencies, we computed Mutual Information (Fig. 3), which consistently confirmed that `danceability` and `loudness` maintain the strongest associations with explicitness. These findings highlight the importance of rhythm-related and acoustic intensity features as key indicators in distinguishing explicit songs.

3.3 Applied Preprocessing

Preprocessing for Classification. Preprocessing was performed on two types of textual representations: traditional TF-IDF vectorization and pre-computed LyricsBERT embeddings, each with its own processing requirements.

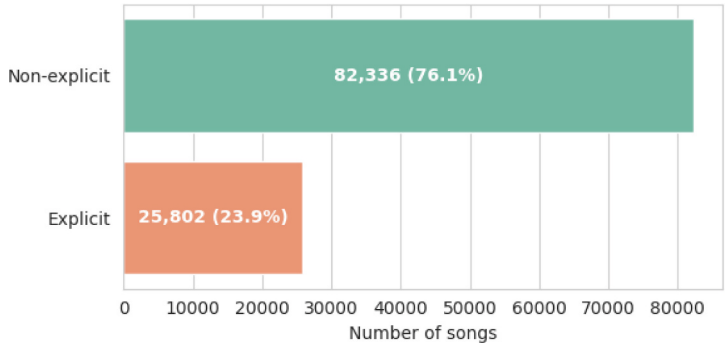


Fig. 1. Distribution of explicit vs non-explicit songs.

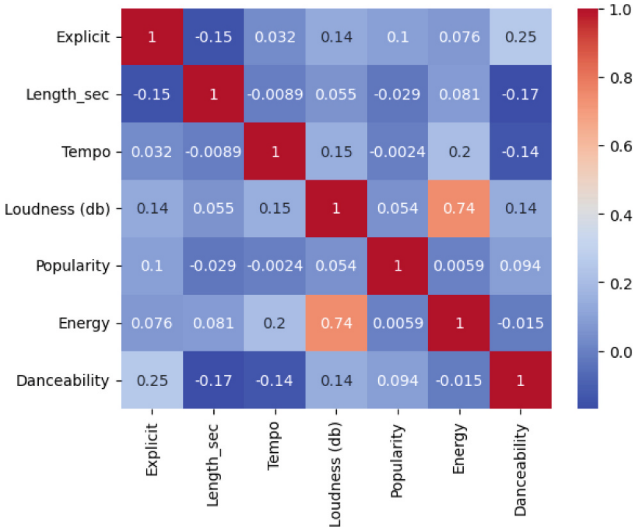


Fig. 2. Correlation matrix between numerical variables.

Textual data drawn from multiple columns underwent a series of transformations, including text cleaning, tokenization, and stemming, followed by TF-IDF vectorization or alternatively, embedding generation with LyricsBERT. Subsequently, these textual features were integrated with normalized numerical attributes.

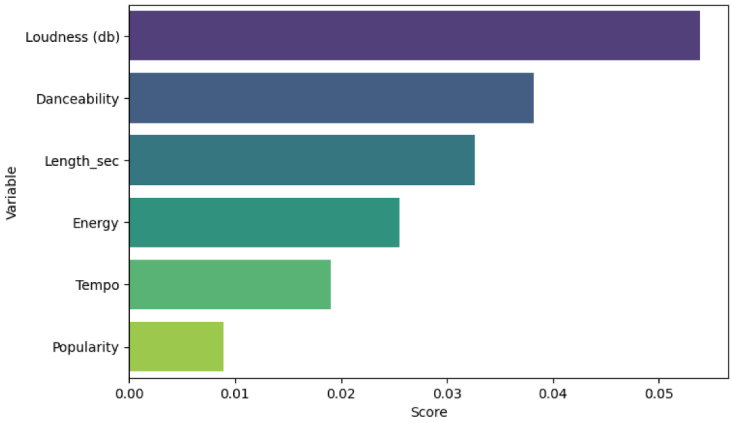


Fig. 3. Mutual Information between acoustic variables and `explicit`.

4 Experimentation and Results

4.1 Hardware and Hyperparameters

To address the task of explicit song detection, we employed TF-IDF and Lyrics-BERT to create embeddings. The first approach was easy to implement, since it is a traditional method and the computation resources were sufficient. In contrast, the second approach was more demanding; for example, processing 1,000 songs required 12 GB of RAM. Therefore, we used an academic account of Khipu (a high-performance computing cluster of the University of Engineering and Technology). Specifically, we used the type of account "investigacion1", which provides 48 CPUs, 40 shards, and 130 GB of memory, with a maximum job time limit of 2 days. Furthermore, hyperparameter tuning was performed using GridSearchCV from the scikit-learn library.

In this work, we use the term *metadata* to refer to all non lyrical features of the dataset. These include both numerical variables (e.g., loudness, energy) and categorical variables (e.g., genre, emotion).

4.2 Comparison Between TF-IDF and LyricsBERT Using Traditional Machine Learning Models

First, we hypothesized that song lyrics would be the primary features for tackling this task of detecting explicit content. The results of this approach are presented in Table 1. Second, to maximize the use of the dataset, we concatenate the lyrics (text features) with the numerical metadata. The results are presented in Table 2.

Table 1. Performance of traditional models using only lyrics (text features) vectorized with TF-IDF and LyricsBERT.

Model	Vectorizer	Accuracy	Precision	Recall	F1-score
Decision Tree	TF-IDF	0.8625	0.6789	0.8043	0.7363
Logistic Regression	TF-IDF	0.8270	0.6102	0.7614	0.6775
Naive Bayes	TF-IDF	0.8213	0.6072	0.7101	0.6546
SVM	TF-IDF	0.6735	0.3845	0.6132	0.4726
Logistic Regression	LyricsBERT	0.8010	0.5610	0.7632	0.6467
Naive Bayes	LyricsBERT	0.6431	0.3779	0.7678	0.5065
Decision Tree	LyricsBERT	0.6271	0.3603	0.7258	0.4815
SVM	LyricsBERT	0.3483	0.2609	0.9450	0.4089

Table 2. Performance of traditional models using lyrics (text features) concatenated with numerical metadata.

Model	Vectorizer	Accuracy	Precision	Recall	F1-score
Decision Tree	TF-IDF	0.8852	0.7201	0.8486	0.7791
Logistic Regression	TF-IDF	0.8454	0.6397	0.8058	0.7132
Naive Bayes	TF-IDF	0.8299	0.6230	0.7273	0.6711
SVM	TF-IDF	0.6336	0.3753	0.8064	0.5122
Logistic Regression	LyricsBERT	0.8368	0.6233	0.7990	0.7003
Decision Tree	LyricsBERT	0.7983	0.5564	0.7624	0.6433
Naive Bayes	LyricsBERT	0.6592	0.3936	0.7921	0.5259
SVM	LyricsBERT	0.3264	0.2548	0.9473	0.4016

Table 3. Performance of traditional models using lyrics (text features) concatenated with numerical and categorical metadata.

Model	Vectorizer	Accuracy	Precision	Recall	F1-score
Decision Tree	TF-IDF	0.8876	0.7132	0.8843	0.7896
Logistic Regression	TF-IDF	0.8888	0.7267	0.8558	0.7860
Naive Bayes	TF-IDF	0.8684	0.7018	0.7795	0.7386
SVM	TF-IDF	0.6501	0.3957	0.8851	0.5469
Logistic Regression	LyricsBERT	0.8391	0.6262	0.8078	0.7055
Decision Tree	LyricsBERT	0.8087	0.5781	0.7337	0.6467
Naive Bayes	LyricsBERT	0.6769	0.4072	0.7773	0.5344
SVM	LyricsBERT	0.3624	0.2623	0.9229	0.4085

Third, we first vectorize the lyrics (text features) together with the categorical metadata, and then concatenate the resulting embeddings with the numerical

Table 4. Performance of traditional models using only metadata (numerical and categorical features), excluding lyrics.

Model	Vectorizer	Accuracy	Precision	Recall	F1-score
Logistic Regression	TF-IDF	0.9351	0.8261	0.9223	0.8715
Naive Bayes	TF-IDF	0.9190	0.8035	0.8746	0.8375
Decision Tree	TF-IDF	0.8749	0.7247	0.7669	0.7452
SVM	TF-IDF	0.8088	0.5610	0.9124	0.6948
Logistic Regression	LyricsBERT	0.8625	0.6707	0.8328	0.7430
Decision Tree	LyricsBERT	0.8298	0.6211	0.7351	0.6733
Naive Bayes	LyricsBERT	0.7683	0.5090	0.8136	0.6262
SVM	LyricsBERT	0.5805	0.3322	0.7506	0.4606

metadata. The results are shown in Table 3. Fourth, we evaluate the performance of traditional models when excluding lyrics and using only metadata (numerical and categorical features). In this experiment, categorical metadata was transformed into embeddings using TF-IDF and LyricsBERT, which were then concatenated with the numerical features. The results are reported in Table 4.

Next, we summarize the top 2 F1-scores obtained across all experimental setups in Fig. 4.

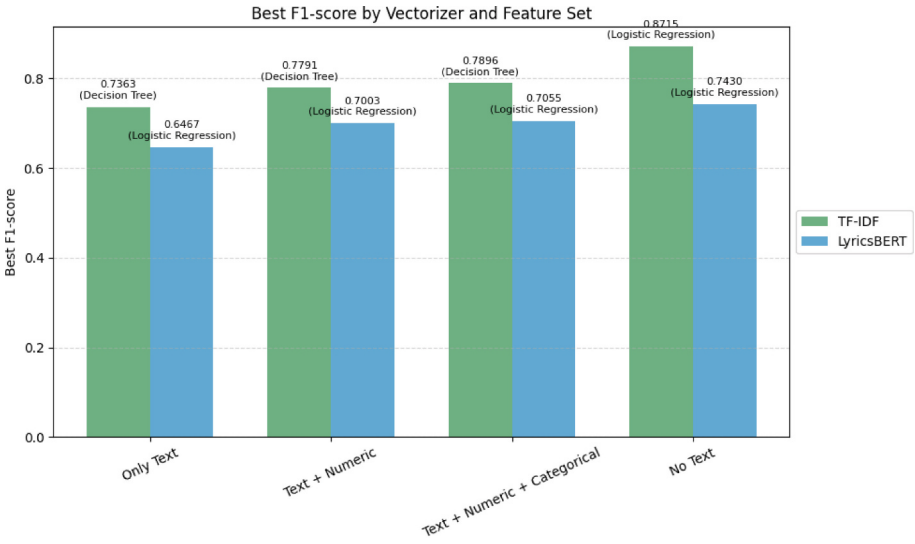


Fig. 4. Comparison of the top 2 F1-scores across different feature combinations and experimental setups.

4.3 BERT Classifier (End-to-End) and MLP with LyricsBERT Embeddings

We used the BERT-base-uncased model, fine-tuned for 3 epochs using the training data, which contained 82,336 instances labeled "No" and 25,802 instances labeled "Yes" for explicit content.

Additionally, we trained a Multilayer Perceptron (MLP) using mean squared error loss. The architecture consists of linear layers interleaved with ReLU activations, except for the last layer, which is linear. The network has an input size of 300 (corresponding to LyricsBERT embeddings), one hidden layer of size 4, and an output layer of size 1. The total number of trainable parameters was 1,209.

The results of both models are summarized in Table 5.

Table 5. Comparison of performance metrics between BERT Classifier and Multilayer Perceptron (MLP)

Metric	BERT Classifier	MLP
Accuracy	0.8638	0.8335
Precision	0.7302	0.7607
Recall	0.6808	0.4521
F1-score	0.7046	0.5671

5 Discussion

The experimental results reveal an interesting trend: traditional models using TF-IDF representations consistently outperform those using LyricsBERT embeddings. Across all four traditional classifiers, the highest F1-score achieved with TF-IDF is greater than that obtained with LyricsBERT. This outcome is likely due to the nature of TF-IDF, which emphasizes word frequency and is particularly effective in capturing explicit terms that act as strong indicators for the classification task. In contrast, LyricsBERT embeddings provide a richer semantic representation but may dilute the impact of these explicit keywords.

A particularly notable finding is that the best overall F1-score was obtained in the "No text" configuration, where only categorical and numerical metadata was used in combination with TF-IDF. This suggests that metadata contains highly predictive information for explicit content detection, sometimes even surpassing the contribution of the lyrics themselves.

Another important aspect is the dimensionality of TF-IDF representations. In our experiments, lyrics-based TF-IDF vectors reached up to 5,000 dimensions. Such high-dimensional and sparse representations may overshadow the contribution of other variables, such as categorical or emotional features, preventing the

model from fully exploiting them. This helps explain why configurations that combine lyrics with metadata did not achieve the highest scores.

Finally, the results highlight a crucial trade-off between computational efficiency and predictive performance. Traditional models with TF-IDF achieved highly competitive results compared to transformer-based approaches, while requiring far fewer computational resources. For instance, the metadata-only model with TF-IDF reached the highest F1-score of 0.8715, compared to 0.7430 for the fine-tuned BERT Classifier and 0.7055 using LyricsBERT embeddings, yet at only a fraction of the computational cost. This efficiency advantage makes TF-IDF based approaches particularly appealing for large-scale applications, such as music streaming platforms, where millions of songs must be processed daily and computational scalability is a primary concern.

6 Conclusions

This research presents a comprehensive evaluation of machine learning approaches for automatic explicit content detection in songs, revealing several significant findings that challenge conventional assumptions in the field.

6.1 Computational Efficiency vs. Performance Trade-Offs

Despite the widespread adoption of transformer-based embeddings, our results show that traditional TF-IDF representations combined with classical models remain highly competitive. They deliver strong predictive performance while requiring far fewer computational resources, making them an attractive and efficient alternative for large-scale applications.

6.2 The Primacy of Musical Metadata over Lyrical Content

A central finding of this study is that metadata only approaches consistently outperformed text-based methods across all evaluated models. Acoustic and categorical features such as danceability, loudness, and genre proved to be more reliable indicators of explicit content than lyrics.

This phenomenon can be attributed to several factors. First, explicit songs often belong to specific musical genres (e.g., rap, hip-hop) that have distinctive acoustic signatures. Second, the production style of explicit content frequently emphasizes certain audio characteristics like increased loudness and danceability to match the energy of the lyrical content. Third, the cultural and contextual associations between musical styles and explicit themes create patterns that are more consistent and easier to detect than the highly variable linguistic expressions found in lyrics.

6.3 Implications for Content Moderation Systems

The results suggest that metadata-based systems offer a more efficient and accurate solution for explicit content detection in streaming platforms. Unlike text-based methods, they enable real-time processing of large catalogs while avoiding challenges such as language dependency, slang, and cultural variability. A metadata-based system would be inherently language-agnostic and less susceptible to linguistic nuances that challenge traditional NLP approaches.

6.4 Feature Engineering and Data Quality Insights

The incorporation of multiple feature types revealed interesting patterns in feature importance. The combination of textual, numerical, and categorical features achieved a high F1-score of 0.7896 with DecisionTree and TF-IDF, representing the highest score for this specific configuration, though higher F1-scores were obtained with other setups. This suggests that while textual information provides some discriminative power, it can also introduce noise that hinders overall classification performance.

The exploratory analysis revealed that danceability and loudness showed the strongest correlation with explicit content, confirming that acoustic properties are reliable predictors. The temporal analysis showing genre evolution across decades also highlights the importance of considering musical trends and cultural context in content classification systems.

6.5 Limitations and Future Directions

While our findings provide valuable insights, several limitations should be acknowledged. The dataset's inherent class imbalance (23.9% explicit content) may influence model performance, and the reliance on Spotify's explicit labels introduces potential noise from inconsistent manual annotation practices. Additionally, our analysis focused on English-language content, and generalization to multilingual datasets requires further investigation.

Future research should explore the applicability of these findings across different languages and cultural contexts. The development of hybrid systems that intelligently combine metadata-based classification with selective lyrical analysis for edge cases could further improve performance. Additionally, investigating the temporal stability of metadata-based models as musical trends evolve would provide insights into system maintenance requirements.

6.6 Practical Recommendations

Based on our findings, we recommend that music streaming platforms prioritize metadata-based classification systems for detecting explicit content. These systems should incorporate acoustic features (danceability, loudness, energy), categorical metadata (genre, emotion), and temporal information (release decade) as

primary predictors. Lyrical analysis should be reserved for cases where metadata-based classification yields uncertain results or for providing explanatory information about the specific nature of explicit content.

For organizations with limited computational resources, TF-IDF-based approaches offer an excellent balance between performance and efficiency. However, for maximum accuracy and scalability, metadata-only approaches using traditional machine learning algorithms represent the optimal solution.

In conclusion, this research demonstrates that effective explicit content detection in songs can be achieved through computationally efficient, metadata-based approaches that outperform more complex text-based methods. These findings not only advance our understanding of music content classification but also provide practical guidance for developing scalable content moderation systems in the music industry.

Acknowledgments. The computational work of this project was supported by the resources of the HPC Khipu cluster (https://web.khipu.utec.edu.pe/) of the University of Engineering and Technology (UTEC).

References

1. Belcik, N.T.: Parental advisory explicit lyrics: a case study of music censorship and suppression in America. In: Young Historians Conference, 10 (2012). https://pdxscholar.library.pdx.edu/younghistorians/2012/oralpres/10
2. Fell, M., Cabrio, E., Corazza, M., Gandon, F.: Comparing automated methods to detect explicit content in song lyrics (2019). https://hal.science/hal-02281137v1
3. Ramos, L., Palma Preciado, C., Kolesnikova, O., Saldaña Pérez, M., Sidorov, G., Shahiki Tash, M.: IntelliLeksika at HOMO-MEX 2024: detection of homophobic content in Spanish lyrics with machine learning. In: IberLEF@SEPLN (2024). https://api.semanticscholar.org/CorpusID:273190406
4. Bergelid, L.: Classification of explicit music content using lyrics and music metadata. Master's thesis, KTH Royal Institute of Technology, School of EECS, TRITA-EECS-EX 2018, 338 (2018). https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-231461
5. Rospocher, M.: Detecting explicit lyrics: a case study in Italian music. Lang. Resour. Eval. **57**(2), 849–867 (2022). https://doi.org/10.1007/s10579-022-09595-3
6. Chin, H., Kim, J., Kim, Y., Shin, J., Yi, M.Y.: Explicit Content Detection in Music Lyrics Using Machine Learning. In: IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 431–434 (2018). https://doi.org/10.1109/BigComp.2018.00085
7. Rospocher, M., Eksir, S.: Assessing fine-grained explicitness of song lyrics. Information **14**(3), 159 (2023). https://doi.org/10.3390/info14030159
8. Chen, X., et al.: A novel approach for explicit song lyrics detection using machine and deep ensemble learning models. PeerJ Comput. Sci. **9**, e1469 (2023). https://doi.org/10.7717/peerj-cs.1469
9. Bolla, B.K., Pattnaik, S.R., Patra, S.: Detection of Objectionable song lyrics using weakly supervised learning and NLP techniques. Procedia Comput. Sci. **235**, 1929–1942 (2024), ICMLDE 2023. https://doi.org/10.1016/j.procs.2024.04.183

10. Dwiyani, L.K.D., Suarjaya, I.M.A.D., Rusjayanthi, N.K.D.: Classification of explicit songs based on lyrics using random forest algorithm. J. Inf. Syst. Inf. **5**(2), 550–567 (2023). https://doi.org/10.51519/journalisi.v5i2.491
11. Podolak, J., Łukasik, S., Balawender, P., Ossowski, J., Bąkowicz, K., Sankowski, P.: Analyzing the influence of language model-generated responses in mitigating hate speech on social media directed at Ukrainian refugees in Poland. arXiv (Cornell Univ.) (2023). https://doi.org/10.48550/arxiv.2311.16905
12. Egivenia, E., Setiawan, G.R., Mintara, S.S., Suhartono, D.: Classification of explicit music content based on lyrics, music metadata, and user annotation. In: SIET'21: 6th International Conference on Sustainable Information Engineering and Technology, pp. 265–270 (2021). https://doi.org/10.1145/3479645.3479695

Spam Detection in YouTube Comments Using a Multinomial Naive Bayes Classifier

Henry Ivan Condori-Alejo, Deyvis Mamani-Lacuta, and Guina Sotomayor Alzamora(✉)

Universidad Nacional del Altiplano, Puno, Peru
{hcondori,gsotomayor}@unap.edu.pe, demamanil@est.unap.edu.pe

Abstract. Spam comments are common on platforms like YouTube and are often used to scam users through spam campaigns masquerading as investment advice from so-called experts, particularly in finance-themed videos. This study aimed to develop and evaluate a Multinomial Naive Bayes model to classify spam comments. For this, the approach used defined rules for classifying spam, created a dataset of over 25,000 comments from 30 finance-related YouTube videos, and generated and evaluated a proposed model using assertiveness metrics to validate its performance. The results showed that the proposed model classified Spanish-language spam comments in finance videos with 98% Recall rate, indicating strong performance based on standard metrics. In conclusion, a viable alternative to reduce spam comments was proposed that could be incorporated into a web tool.

Keywords: spam comments · Multinomial Naive Bayes · NLP · CRISP-DM · YouTube

1 Introduction

YouTube is a popular platform that allows users to upload, share, and comment on published videos, among other features. Its main feature is the large amount of videos covering a wide range of topics. The platform is available in over 60 languages and many countries around the world. It is also one of the most widely used platforms globally, with more than 1.5 billion monthly active users. On average, users spend about one hour per day watching videos, considering only mobile devices [1,2].

In 2015, YouTube implemented a monetization system in order to reward content creators for their work and encourage continued platform engagement [3]. To monetize a channel, it is necessary to have more than 1,000 subscribers and 400 h of watch time in the last 12 months. This change led many smaller channels to post spam comments promoting their content on high-traffic videos [4] since the presence of spam comments generates a negative user experience and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 207–221, 2026.
https://doi.org/10.1007/978-3-032-20322-9_15

also increases the risk of users being scammed. Consequently, content creators manually review and remove spam comments [5].

Today, this situation has escalated into a critical problem, as cybercriminals exploit the YouTube comment section as a gateway to reach potential victims. They employ increasingly sophisticated techniques such as bots to carry out large-scale spam campaigns [6]. One common tactic involves impersonating the video creator by copying their username and profile picture to gain trust and increase the likelihood of success [7].

These spam campaigns occur most frequently in finance-related videos, as this type of content typically attracts users who are more interested in investing or generating income [8]. The campaigns involve posting a high volume of spam comments recommending so-called "investment experts," using fake user accounts to encourage viewers to contact them via external channels such as WhatsApp, Instagram, or Telegram, where the scam is ultimately carried out [9].

Spam comments have a direct impact on content creators, as they degrade the user experience for their audience and followers. In some cases, creators agree to disable the comment section [10]. However, that deprives them of valuable feedback on the shared video, such as constructive criticism or ideas for future content, essentially breaks the main channel of interaction with their audience [1,10].

The process of removing spam comments is an arduous task that involves searching, identifying, and eliminating such comments, which can consume significant time and effort from content creators depending on the volume of comments and video views. On average, it can take around twenty minutes per day per a video with approximately 200,000 views [5].

Although YouTube has its built-in spam filtering system, it appears to be ineffective [4], since a large number of spam comments continuously appear. According to Gandra, as cited in Samsudin [11], on average, one out of every hundred comments is spam. Therefore, this study aims to develop a Multinomial Naive Bayes model to classify spam comments in Spanish for finance-related YouTube videos.

There are researchers using machine learning on spam comments in music-related YouTube videos like Oh [4] with the Bernoulli Naive Bayes algorithms; Ifriza & Saman [12] with the Gaussian Naive Bayes algorithm, and Samsudin et al. [11] with the classic Naive Bayes algorithm. These approaches use standard metrics evaluation, achieving accuracy scores above 80%. Additionally, Narvaés-Albuja [13] uses a Multimodal Naive Bayes algorithm for spam comments in YouTube finance videos, reaching a 74% F1 score, which is lower than the 80% score achieved in this work. Other classification techniques used for spam detection in YouTube comments include Naive Bayes, Random Forest, and Support Vector Machines [14].

The main contribution of this study is the development of a web-oriented Multinomial Naive Bayes model for classifying Spanish-language spam comments in finance-related YouTube videos. Furthermore, this work provides a manually labeled, up-to-date dataset of finance video comments, containing over 20,000

records collected from thirty finance videos published between 2022 and 2023. In the future, this model could be directly integrated into a web-based tool, eliminating the need for third-party services, such as a dedicated VPS. The generated dataset may also serve as a basis for future research.

The structure of this work is as follows: Sect. 1 presents the introduction; Sect. 2 reviews the state of the art; Sect. 3 outlines the methodology; Sect. 4 presents the results; Sect. 5 presents the discussions; Sect. 6 offers the conclusions; and finally, the references are listed.

2 State of the Art

2.1 YouTube and Its Ecosystem

YouTube is one of the largest and most popular video platforms in the world, with over 2 billion monthly active users. Founded in 2005 and acquired by Google in 2006, YouTube has become a major audiovisual content search engine, hosting videos across a wide variety of categories, from entertainment to education. Users can upload, share, comment, and interact with videos, creating a dynamic ecosystem where viewer feedback—often expressed in the comment section—plays a crucial role in the overall user experience [15].

The YouTube comment section is a key space for interaction between creators and viewers. However, this very characteristic also makes it an attractive target for malicious activities, such as the posting of spam comments, which can degrade the quality of interaction and, in more severe cases, lead users to fall victim to scams.

2.2 Spam Comments in Finance Videos

Spam comments are unwanted, often automated messages posted on online platforms to deceive, scam, or aggressively promote products and services. On YouTube, these comments are common in finance-themed videos, where fraudulent schemes promise high returns with little or no risk. These comments include links to external sites or contact information to continue the scam outside the platform [6,9].

The rise of spam in YouTube video comments generate automated and efficient methods for its detection and removal. Although YouTube implements spam filtering systems, these are not fully effective, especially against advanced techniques that use bots to generate and post comments at scale [7].

2.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of artificial intelligence focused on the interaction between computers and human language. Its main goal is to enable machines to understand, interpret, and generate human language in a meaningful way. NLP applications are diverse and include machine translation, sentiment analysis, and text classification [16].

Within NLP, one of the most common tasks is text classification, which involves assigning a label or category to a piece of text based on its content. In the case of spam detection, text classification enables the differentiation between legitimate and spam comments using models trained on large, labeled datasets.

2.4 Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a classification method based on Bayes' probability theory. It is widely used for text classification due to its simplicity and efficiency. The Naive Bayes model assumes that features (in this case, words) are independent of one another, which simplifies the computation of the probability that a comment belongs to a specific category (spam or not spam) [17].

This model is particularly effective in scenarios involving the classification of text into multiple categories, such as detecting spam in YouTube comments. The Algorithm 1 and 2 calculates the probability that a comment is spam based on the frequency of words present in the text and their association with predefined categories [21].

Algorithm 1 TrainMultinomialINB(C,D)

$V \leftarrow$ ExtractVocabulary(D)
$N \leftarrow$ CountDocs(D)
for each $c \in C$ **do** $N_c \leftarrow$ CountDocsInClass(D,c)
    prior[c] $\leftarrow \frac{N_c}{N}$
    $\text{text}_c \leftarrow$ ConcatenateTextOfAllDocsInClass(D,c)
    for each $t \in V$ **do** $T_{ct} \leftarrow$CountTokensOfTerm(text_c,t)
    end for
    for each $t \in V$ **do** condprob[t][c] $\leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
    end for
end for
return V, prior, condprob

Algorithm 2 ApplyMultinomialNB(C,V,prior,condprob,d)

$W \leftarrow$ ExtractTokensFromDoc(D)
for each $c \in C$ **do** score[c] $\leftarrow$log prior[c]
    for each $t \in W$ **do** score[c]+= log condprob[t][c]
    end for
end for
return arg $\max_c \in C$ score[c]

3 Methodology

The research approach is quantitative and experimental, as variables will be manipulated and their effects observed in a controlled environment. This study used Spanish-language comments posted on finance-related YouTube videos, with a sample consisting of more than 25,000 comments extracted from 30 selected videos in this category. The content was recent, published between 2022 and 2023, and attracted a high volume of user interaction.

In this context the identification of spam in comments on finance videos negatively affects both content creators and viewers, who may become victims of fraud. The model must efficiently classify spam comments, thereby improving the user experience on YouTube. The following phases are proposed:

1. **Classification Rules definition**. This phase, analyzes the comment distribution, studies the word frequency, and identifies relevant patterns that could influence classification.
2. **Dataset Construction**. This phase involves the initial collection and exploration of data. It is a critical stage where raw data are filtrated and cleaned, and transformed to be suitable for machine learning models. This process removes duplicate and empty comments, removes Special characters and stopwords, and applies stemming to reduce words to their root form using a text preprocessor. Then, use the Bag of Words technique to convert text into numerical vectors that the algorithm can process using a text formating. Finally, use outliers and imputations in the preprocessing.
3. **Modeling**. This phase uses the spam classifier model through the Multinomial Naive Bayes algorithm, which is suitable for text classification where features are word frequencies. The trained model uses the prepared dataset and its adjusted parameters to optimize performance to indicate if video comments are classified as SPAM or not SPAM, using a confusion matrix.
4. **Evaluation**. The standard metrics evaluate the model, such as Precision, Recall, and F1 score.

4 Results

This section presents and analyzes the results obtained from the implementation of the Multinomial Naive Bayes model for classifying Spanish-language spam comments in finance-related YouTube videos. In this work, the proposed architecture is shown in Fig. 1, involving Classification Rules Definition, Dataset Construction, Modeling, and Evaluation to determine its assertiveness in identifying spam comments. Additionally, the implications of the discussed results are in the context of content moderation on online video platforms.

The experiments were conducted on an ASUS FX504GE laptop, equipped with an Intel® Core™i5-8300H CPU @ 2.30 GHz, 8 GB RAM, 931 GB Seagate ST1000LM035-1RK172 (SATA) storage, running Windows 10 Pro 64-bit, with a Generic PnP Monitor (1920×1080@60 Hz). The software used included Visual

Studio Code version 1.8, Node JS version 18.12.0 (LTS), JavaScript, a spreadsheet tool, and Brave Browser version 1.52.130. Online tools included Google Colaboratory using Python 3.10.12. Additionally, data resources were obtained through the free YouTube Data API v3, accessed through tokens managed via the Google Console.

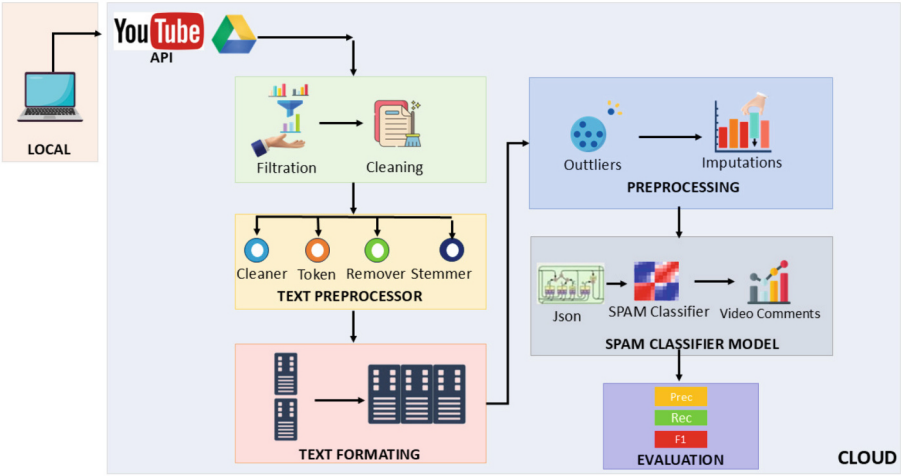


Fig. 1. Proposed processing architecture.

4.1 Classification Rules Definition

To classify comments, a set of rules was defined based on the frequency of words and phrases commonly associated with spam in finance-related videos, to filter and clean the collected data, similar to the approach in Brown [9]. These rules included the identification of keywords such as guaranteed profit ("ganancia garantizada"), safe investment ("inversión segura"), contact via WhatsApp ("contacto por WhatsApp"), "Telegram", and "Instagram", which are typically used in fraudulent comments to lure users into scam schemes. Other phrases referenced requests for investment experts, assistance from experts, recommendations, trading and cryptocurrency advice, and mentions of results or profits obtained through experts.

It is important to note that since YouTube uses unique usernames, cybercriminals can no longer impersonate well-known channel names. Therefore, this study considers only the textual content of the comment and does not analyze usernames. Figure 2 and Fig. 3 show examples of the most frequently occurring words in spam and non-spam comments, respectively. These figures clearly illustrate how certain words and phrases dominate in spam versus non-spam comments, justifying their use in rule-based classification.

Fig. 2. Most frequent words in spam comments.

Fig. 3. Most frequent words in non-spam comments.

4.2 Dataset Construction

This work analyzes YouTube, one of the most-used platforms globally, although still vulnerable to cybercriminals conducting spam campaigns. Despite its internal spam filtering, YouTube's system appears inefficient [1,4], as a significant number of spam comments persist. On average, one out of every hundred comments is spam [11].

The situation assessment identified key requirements: access to a Google account for services such as Drive, Colaboratory, and the Developer Console, as well as access to YouTube Data API credentials to extract data. The dataset uses the API's daily token quota to fetch all comments and replies from selected videos without data loss. Comments were removed or reported, and no paid plan was required to complete the research.

Project constraints included limiting the analysis to Spanish-language comments from finance-related YouTube videos. The dataset size depended on the time available for its development and the video selection criteria. Finally, assess the financial and data-related risk and define the contingency plans. And the costs identified are that the manual labeling of the dataset required significant effort, and the entire development and implementation took two months [19,20]. On the other hand, the benefits included reducing spam volume, lowering the

time required to report/delete spam, giving content creators more control, and enhancing the user experience in comment sections.

For initial data collection, generate an API key by YouTube Data API. After activation, by creating credentials, grant access to up to 10,000 daily requests. Select finance-related videos through standard search queries using terms like finance videos ("videos de finanza"). Filter results with criteria as publication between 2022–2023, financial theme, Spanish-language comments, and visible indicators of spam.

The extracted comments generate a dataset with the following fields: video identifier (video_id), comment identifier (comment_id), author name (author), and comment content (comment). Where no includes metadata such as posting date or like count. Finally, the resulting dataset contained 25,672 Spanish-language comments from 30 finance videos, as shown in Fig. 4. Most selected videos had fewer than 1,000 comments; the smallest had 152 and the largest had 2,840 comments.

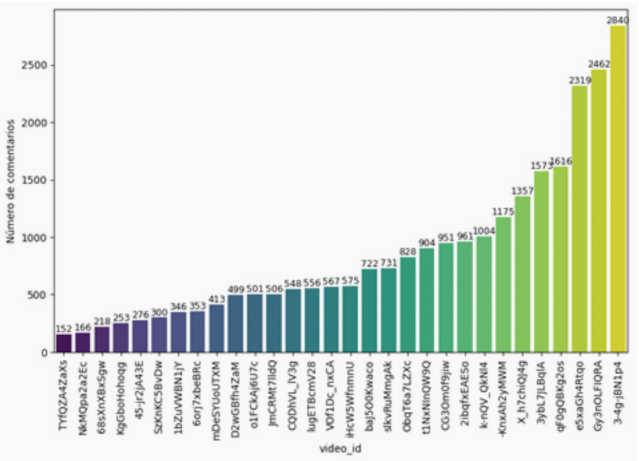


Fig. 4. Number of comments per video.

All comments were manually labeled as spam or not spam and stored in the class field ("class"), as shown in Fig. 5. The balanced dataset produced is suitable for training and evaluating the Multinomial Naive Bayes model.

The Data cleaning and preparation ensure the dataset's quality and representativeness. For this purpose, data exploration focused on the "comment" attribute. This work uses word clouds and frequency tables, excluding single characters, symbols, and numbers. They were treated as natural stopwords due to their lack of semantic significance. Additionally, remove the 3,709 duplicate comments across both classes and all empty comments (3) to avoid bias in the final results and their interpretation. The final dataset consisted of 21,963 records,

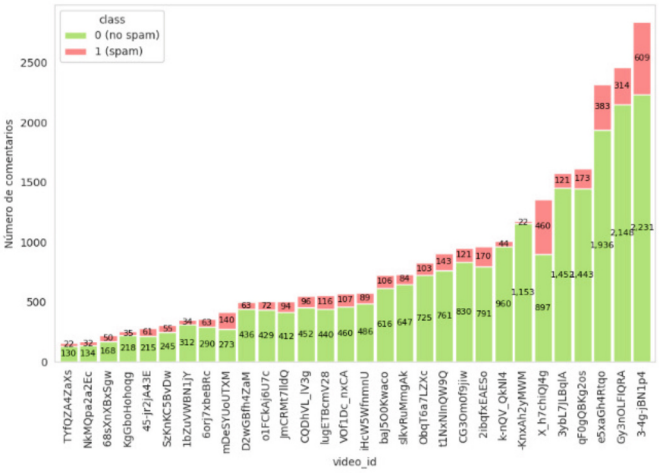


Fig. 5. Number of comments spam and no spam per video.

with 19,298 non-spam and 2,665 spam comments. Due to this imbalance, it resampled data to equalize class representation.

The most frequent words in non-spam comments included “de,” “que,” “en,” “la,” “el,” “no,” “es,” “para,” “un,” and “me,” with frequencies ranging from 15,941 to 4,826. Similarly, the most frequent words in spam comments included “de,” “la,” “con,” “que,” “en,” “es,” “el,” “ella,” “mi,” and “un,” with frequencies from 4,909 to 1,247. In total, 145 words have high frequency across both classes and were added to the stopword list. This filter allows the clear visualization of representative non-spam terms, which helps distinguish between spam and non-spam comments.

Figure 6 shows the most frequent words within non-spam comments, which are part of the stopwords. These words calculate the probability that a comment belongs to the non-spam class or not, based on the frequency of its component words. Similarly, Fig. 7 shows the frequent words within spam comments. Here, the most frequent words correspond to articles such as “ella” and “él”, or “sra”. This makes sense since the pattern of referring to “experts” using articles instead of their direct names is how each of the bot accounts replies to the main comments.

The data quality verification ensured the completeness dataset and included the necessary attributes to achieve the project’s objectives. The dataset was determined to be valid and free of high-risk errors, further corrections were simple.

4.3 Classification Model Generation

Data integration followed the activity flow shown in Fig. 8. The dataset was split into two subsets: 80% for training and 20% for testing. Then, applies the random

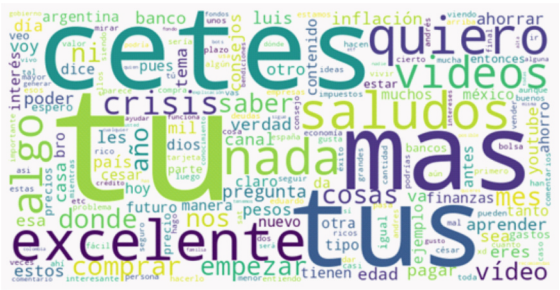


Fig. 6. Most representative words in non-spam comments.

Fig. 7. Most representative words in spam comments.

oversampling to the training set, storing it in JSON format with "comment" and "class" keys. Afterward, the process transitioned from Google Colaboratory to the Node.js environment.

For data formatting, a custom "Text Preprocessor" was implemented. This component accepted a text string ("comment") and returned a token array. The functions included:

1. **Cleaner**. This method receives a text string and returns a clean text string, changing the entire string to lowercase and removing several character types, such as: at, quotes, punctuation marks, numbers, spaces, and so on.
2. **Tokenizer**. This method receives a text string and returns an array of words (tokens). It performs the separation of the string, considering the intervening white spaces between words.
3. **StopWordsRemover**. This method takes a token array (words), evaluates each one, and discards any tokens that fall within the stopwords list.
4. **Stemmer**. This method takes a token array and returns another array containing the result of passing the input tokens through the PorterStemmer stemmer. In general, PorterStemmer converts a word to its base form.
5. **Preprocessor**. This method receives a text string as input and returns a token array as a result by sequentially applying each method described above.

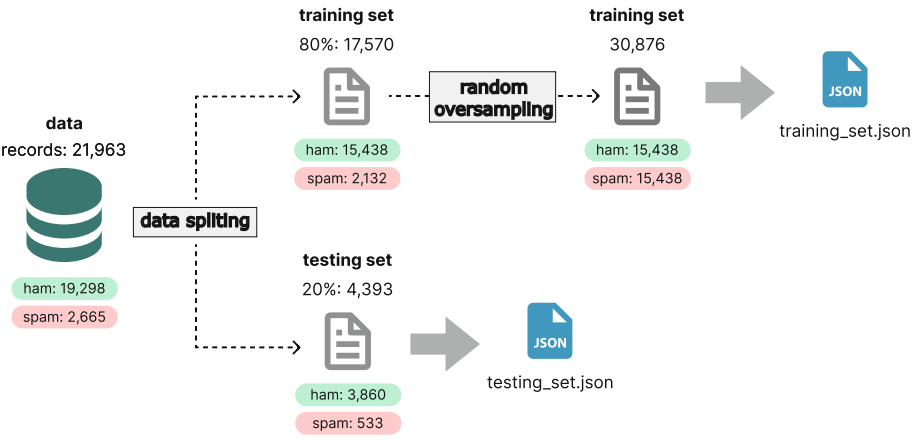


Fig. 8. Data integration procedure.

Both training and testing datasets were formatted to match the input expected by the Multinomial Naive Bayes classifier, as illustrated in Fig. 9. While the "Renaming Keys" function adjusted JSON key names to follow conventional naming standards.

The Multinomial Naive Bayes model based on the algorithm proposed by Manning et al. [21] was implemented in JavaScript. Changes included setting class priors to zero and applying a proportion threshold of 2.5 for classification decisions. This adjustment yielded improved results even from a threshold of 1.0.

The model trains the prepared dataset. During training, fine-tuned parameters optimize the classification performance. Moreover, use the Bag of Words technique to convert text into feature vectors, allowing the model to estimate the spam probability of a comment based on word frequency. As noted, set the prior probability in the Naive Bayes algorithm to zero and the prediction mechanism to return a class (0 = not spam, 1 = spam) based on the class with the highest score, even with a minimal score difference. This behavior increased the rate of false positives, where legitimate comments might be misclassified as spam.

4.4 Model Performance Evaluation

The evaluation of the model uses a test dataset (20%). Since this is a supervised learning task, the specific classification is appropriate to use evaluation metrics derived from the confusion matrix. These are summarized in Table 1, and include Precision, Recall, and F1 Score.

The model achieved a Recall of 98%, indicating that the vast majority of its predictions were correct. In addition, the Precision score and F1 Score were 0.80%, demonstrating the model's ability to identify correctly most spam comments in the test set. These results are summarized in Table 2, which presents evaluation metrics across different scenarios. This table offers a detailed view of

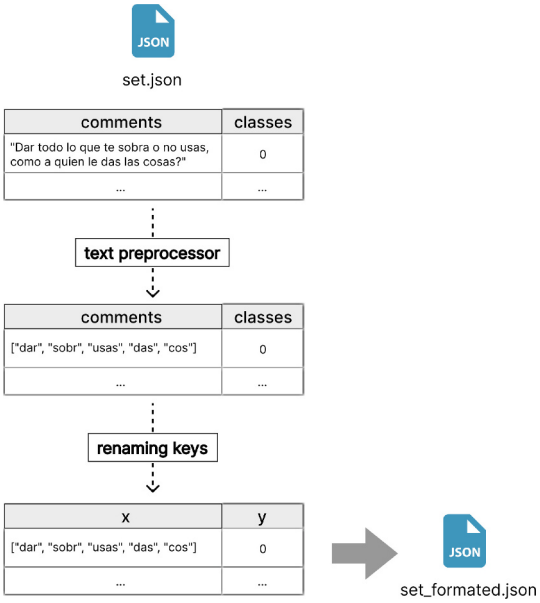


Fig. 9. Dataset formatting procedure.

Table 1. Confussion Matriz Results

		Predicted class		
		SPAM (1)	NO SPAM (0)	TOTAL
Real class	SPAM (1)	TP=437	FN=96	533
	NO SPAM (0)	FP=122	TN=3738	3860
TOTAL		559	3834	4393

Table 2. Metric Results of Umbral Evaluation

METRIC	UMBRAL VALUE						
	1.0	1.3	1.6	1.9	2.2	2.5	2.8
Precision	0.49	0.59	0.66	0.71	0.74	0.78	0.80
Recall	0.98	0.96	0.93	0.89	0.85	0.81	0.77
F1 Score	0.65	0.73	0.77	0.79	0.79	0.80	0.78

model performance under various thresholds, reinforcing the model's robustness in spam classification.

5 Discussions

This paper considered works such as the UCI Machine Learning dataset of music video spam comments from YouTube, used by Oh [4], Ifriza & Sam'an [12], and Samsudin et al. [11], and a financial video spam comments from YouTube, used by Narváez-Albuja [13]. The comparison is summarized in Table 3.

Table 3. Comparison Results

Author	Algorithm	Evaluation Metric	Score	Dataset
Proposal model	Naive Bayes Multinomial	Recall	98%	25,672
Oh [4]	Bernoulli Naive Bayes	F1 Score	81%	2,000
Ifriza & Sam'an [12]	Gaussian Naive Bayes	F1 Score	84%	2,000
Samsudin et al. [11]	Naive Bayes	Precision	87%	2,000
Narváez-Albuja [13]	Naive Bayes Multinomial	F1 Score	74%	255,109

6 Conclusions

The Multinomial Naive Bayes model proved to be highly effective for the task of classifying spam comments in finance-related YouTube videos, achieving a Recall of 98% and a Precision and F1 Score of 80%. These results indicate the model's notable ability to correctly identify spam comments while minimizing both false positives and false negatives.

The dataset quality and the data preparation process were fundamental to the model's success. Removing empty records, applying text cleaning, and properly vectorizing the comments enabled the model to operate optimally. Doing rigorous preprocessing is vital in data mining and machine learning projects, especially when working with unstructured data such as text comments.

The results of this study have important practical implications. The developed model can be integrated into moderation tools for YouTube and other online content platforms, offering an effective solution for detecting and removing spam, improving the user experience, and also helping protect users from potential fraud, particularly in finance-related videos, where the risk of scams is considerable.

Acknowledgments. The authors thank the Universidad Nacional del Altiplano for its support in the development of this research work.

References

1. Abdullah, A.O., Ali, M.A., Karabatak, M., Sengur, A.: A comparative analysis of common YouTube comment spam filtering techniques. In: 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–5 (2018). https://doi.org/10.1109/ISDFS.2018.8355315
2. Binmahboob, T.: YouTube as a learning tool to improve students' speaking skills as perceived by efl teachers in secondary school. Int. J. Appl. Linguist. Engl. Lit. **9**(6), 13–22 (2020). https://doi.org/10.7575/aiac.ijalel.v.9n.6p.13.i2.pp1046-1053
3. Alberto, T.C. , Lochter, J.V., Almeida, T.: TubeSpam: comment spam filtering on YouTube. In: IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 138–143 (2015). https://doi.org/10.1109/ICMLA.2015.37
4. Oh, H.: A YouTube spam comments detection scheme using cascaded ensemble machine learning model. IEEE Access **9**, 144121–144128 (2021). https://doi.org/10.1109/ACCESS.2021.3121508
5. ThioJoe: I Created an App That DESTROYS Scam Comments (Because YouTube Wouldn't)—OUTDATED, READ DESCRIPTION, [Video recording] (2021). https://www.youtube.com/watch?v=-vOakOgYLUI
6. Aiyar, S., Shetty, N.P.: N-Gram Assisted Youtube Spam Comment Detection, vol. 132, pp. 174–182 (2018). https://doi.org/10.1016/j.procs.2018.05.181
7. Vincent, J.: YouTube is cracking down on tricks that spammers use to impersonate creators. The Verge (2022). https://www.theverge.com/2022/6/30/23189367/youtube-spam-account-policy-subscriber-count-special-characters
8. Nesbo, E.: What is a YouTube Spambot? How Can You Stop Them? MUO (2021). https://www.makeuseof.com/what-is-a-youtube-spambot/
9. Brown, A.: YouTube superstars issue new scam warning, how to avoid the trick. Express.Co.Uk. (2022). https://www.express.co.uk/life-style/science-technology/1590429/youtube-comment-spam-scam-warning-trick-trying-to-steal-your-money
10. Alias, N., Foozy, C.F.M., Ramli, S.N.: Video spam comment features selection using machine learning techniques. Indonesian J. Elect. Eng. Comput. Sci. **15**(2), pp. 1046–1053 (2019). https://doi.org/10.11591/ijeecs.v15
11. Samsudin, N.M., Mohd Foozy C.F.B., Alias, N., Shamala, P., Othman, N.F., Wan Din, W.I.S.: Youtube spam detection framework using naïve bayes and logistic regression. Indonesian J. Elect. Eng. Comput. Sci. **14**(3), 1508–1517 (2019). https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517
12. Ifriza, Y.N., Sam'an, M.: Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection. J. Soft Comput. Explor. **2**(2), 2 (2021). https://doi.org/10.52465/joscex.v2i2.42
13. Narváez-Albuja, A., Aplicación antispam de comentarios para prevenir delitos informáticos en la plataforma de YouTube. Tesis de Maestría. (2022). https://reunir.unir.net/handle/123456789/14020
14. Sinhal, A., Maheshwari, M.: An extensive review on contemporary analysis of comment filtration of youtube videos using machine learning techniques. Int. J. Emerging Technol. Adv. Eng. **12**(9), 130–141 (2022). https://doi.org/10.46338/ijetae0922_14
15. Cho, S.E., Suh, J.: Translating korean beauty YouTube Channels for a Global Audience. In: Desjardins, R., Larsonneur, C., Lacour, P. (eds.) When Translation Goes Digital: Case Studies and Critical Reflections, pp. 173–197. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-51761-8_8

16. Vangara, V., Vangara, S.: A survey on natural language processing in context with machine learning. Int. J. Anal. Exp. Modal Anal. **12**, 1390–1395 (2020)
17. Jurafsky, D., Martin, J. H.: Naive Bayes and Sentiment Classification. In: Speech and Language Processing, 3rd. ed. p. 23 (2023). https://web.stanford.edu/~jurafsky/slp3/
18. Wirth, R.,Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, vol. 1, pp. 29–39 (2020)
19. Cloudfactory.: Data Labeling for ML in 2024: A Comprehensive Guide (2024). https://www.cloudfactory.com/data-labeling-guide
20. Ontop.: What is the minimum wage in Peru? (2024). https://www.getontop.com/blog/what-is-the-minimum-wage-in-peru-in-2024
21. Manning, C.D., Raghavan, P., Schütze, H.: Text classification and Naive Bayes. En Introduction to Information Retrieval. p. 35 (2008). https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

Prompt Engineering in Spanish for Improving Sexism Detection in Tweets

María Isabel Limaylla-Lunarejo(✉)

Fac. Informática, Database Lab, Universidade da Coruña, Coruña, Spain
maria.limaylla@udc.es

Abstract. Social media currently plays a significant role in interpersonal interaction and information consumption. However, these platforms have also become spaces where various social issues, such as sexism and gender gaps, are increasingly disseminated. This study evaluates the performance of two large language models, GPT-4o and Mistral 7B, in detecting sexist content on the social media platform X (formerly Twitter), using several prompt combinations on the EXIST 2024 dataset. Our findings show that GPT-4o delivers more consistent results across different prompts and achieves higher f1-scores values. In contrast, Mistral 7B demonstrates greater variability but achieves higher recall in several configurations, allowing it to detect more sexist tweets at the cost of increased false positives. When exploring the impact of prompt design (zero-shot vs. few-shot), we found that including examples in few-shot prompts did not significantly improve performance for GPT-4o and reduced performance for Mistral 7B.

Keywords: sexism identification · prompt engineering · LLM

1 Introduction

Sexism, defined as discrimination or prejudice based on gender, remains a pressing social issue that undermines gender equality and disproportionately affects women. Moreover, sexist content has become increasingly prevalent across the internet. X[1], formerly Twitter, is a highly popular social media platform with over 436 million monthly active users worldwide, originally launched in 2006 [9]. Various types of offensive messages can be found on this social network. For example, the following is considered a sexist tweet: "Hahaha I'm not competing with you, feminazi!!!". The use of Artificial Intelligence (AI) techniques to automatically identify any indication of sexism in tweets has been recommended as a way to detect and respond to this type of violence against women, whether by deleting the tweet, issuing warnings, or suspending accounts [17].

Several AI-based models have been developed for automatically classifying sexism in English-language tweets. However, significantly fewer studies addressing Spanish-language tweets are available in the literature. Rodriguez-Sanchez

[1] https://x.com/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 222–233, 2026.
https://doi.org/10.1007/978-3-032-20322-9_16

et al. [24] present one of the first automatic classification analyses and introduce one of the earliest datasets focused on Spanish. Even though research in this area has increased in recent years, a significant gap remains in the automatic detection of sexist content in Spanish-language tweets.

With the rise of large language models (LLMs), prompt engineering is essential to take advantage of their capabilities. Prompt engineering refers to the process of designing specific instructions, known as prompts, to guide a model toward producing the desired output for a given task. A prompt is a set of instructions given to an LLM to obtain some output or responses. These prompts may consist of sentences, questions, or contextual cues designed to help the LLM produce a targeted result. The importance of prompt design lies in its ability to influence the quality and relevance of a model's responses [25]. According to [13], the prompt itself typically has a greater impact on the output than other parameters, such as temperature. In recent years, prompt-based approaches have been increasingly applied to text classification tasks, particularly in zero-shot and few-shot learning scenarios [12].

In this study, we investigate the impact of using two LLMs with Spanish language support and different prompt formulations on the performance of sexism detection in tweets. We designed several prompts in Spanish and evaluated their effectiveness using the EXIST 2024 dataset to assess model performance. The main contribution is evaluating how Spanish prompt engineering influences classification performance.

The paper is organized as follows. Section 2 presents the related works on sexism identification in tweets. Section 3 presents the research questions, the datasets, and the experimental procedure. Section 4 reports the experiments carried out and their results. Section 5 discusses the results from the previous section. Section 6 presents some relevant threats to validity. Finally, we conclude the paper in Sect. 7.

2 Related Works

Several studies have addressed the automatic detection of sexism in tweets written in English [8,11,15,26]. In the case of the Spanish language, only in recent years have a few studies emerged. In the systematic review by [11], Spanish was the second most commonly used language, with 10 datasets compared to 22 for English. Rodriguez et al. [24] present one of the first analyses of automatic classification of tweets and introduce one of the earliest Spanish-language datasets (MeTwo). The study also highlights challenges related to irony, word ambiguity, limited contextual information, and the relatively small size of the dataset.

In 2021, the EXIST: sEXism Identification in Social neTworks (EXIST 2021) challenge was organized. This challenge was presented as part of the IberLEF 2021 workshop, which aimed to encourage the research community to develop improved NLP tasks in Spanish and other Iberian languages. Since 2021, the EXIST challenge has been held annually. An overview of the 2024 edition is presented in [19]. In this edition, several studies focused on the identification

of sexism (Task 1), employing a range of models. For instance, Maqbool [14] used two traditional machine learning algorithms, Random Forest and XGBoost, as well as the Multilingual BERT (mBERT), with mBERT outperforming the others in identifying sexism in Spanish tweets. Spanish-trained models have been consistently tested in this challenge in recent years. For instance, BETO and MarIA, both BERT-based models [4] trained exclusively on Spanish corpora, were utilized by [16]. Additionally, BETO was employed by [28] for the same task. Additionally, RoBERTuito [20], a Roberta-based model, trained on 500 million Spanish tweets, has also been used in this context [2,32].

Finally, some recent studies have begun to explore the use of prompting and LLMs for this task. Azadi et al. [3] employed few-shot prompting with GPT-3.5; however, in the sexism identification task, fine-tuned XLM-RoBERTa achieved better results, placing 4th overall. Smith et al. [30] also used prompting in English with GPT-4 to perform classification in both English and Spanish, demonstrating strong performance. Similarly, Siino and Tinnirello [29] applied prompt engineering using Mistral 7B with a few-shot approach, translating Spanish inputs into English before classification. English prompts with the Gemini API to address both Spanish and English texts were used in [31], while Quan and Thin [21] evaluated prompts using LLaMA 2 and 3 models.

Despite the growing interest in prompt-based approaches, few studies have explored prompt design in Spanish or examined variations in Spanish-language prompting. This paper aims to fill that gap by examining the impact of prompt formulation on performance for sexism identification.

3 Research Design

This section first introduces the research questions and metrics, followed by a description of the datasets used in this study, a prompt design and models, and an overview of the complete process.

3.1 Research Questions and Metrics

These are the research questions addressed in this study:

- **RQ1**: How does prompt formulation in Spanish influence the performance of LLMs in detecting sexist language?
- **RQ2**: Do few-shot prompts with context and examples outperform zero-shot prompts in Spanish?

To evaluate model performance across different prompts, the following metrics will be used: precision, recall, and f1-score. Precision reflects the proportion of accurate predictions of all predicted classifications. Recall measures how many actual positives were correctly identified. The f1-score combines precision and recall by calculating their harmonic mean [6]. To verify if there is no statistically significant difference in performance, we conducted a multiple-hypothesis testing [22] using a Cochran Q (omnibus test) and the McNemar statistic (post hoc

tests). We considered a significance level (α) of 0.05 for Cochran's Q test and applied a Bonferroni correction [27] for the post hoc test. We used the library MLxtend that implements both tests in Python.

3.2 Dataset

In this study, we utilized the EXIST (sEXism Identification in Social neTworks) dataset from the 2024 event[2]. The EXIST 2024 dataset [19] includes a wide range of sexist expressions and related phenomena, including descriptive or reported statements where the sexist message is conveyed through the narration or description of sexist behavior. The dataset is divided into three splits: training, development (dev), and testing. Labels are provided for the training and development sets, while the test set remains unlabeled for evaluation purposes.

The EXIST training dataset contains a rich variety of popular terms and expressions, comprising 3260 tweets in English and 3660 in Spanish. In contrast to previous years, this dataset also includes detailed information about both the annotators and the labels assigned by each individual annotator. The available annotator metadata includes gender, ethnicity, country of origin, and educational background. We categorize the labels into five groups based on the labels of the six annotators. The distribution of Spanish-language tweets across these groups is shown in Table 1. The dataset contains a relatively balanced distribution between all categories. This study considers only the "Sexist" and "Non-sexist" categories, comprising a total of 1159 tweets.

Table 1. EXIST 2024

Categories	Description	Number
Sexist	All labels are YES	493
Mostly Sexist	Majority of labels are YES	1067
Non-sexist	All labels are NO	666
Mostly Non-sexist	Majority of labels are NO	968
Doubtful	Half of the labels are YES and half are NO	466
Total		3660

3.3 Prompt Design

Our experimental design employed several prompt templates, with initial agent interactions conducted in Spanish. We developed four system prompts and two user prompts. System prompts were used to initialize the API interaction, providing essential definitions and instructions. We explored four distinct approaches:

[2] https://nlp.uned.es/exist2024/.

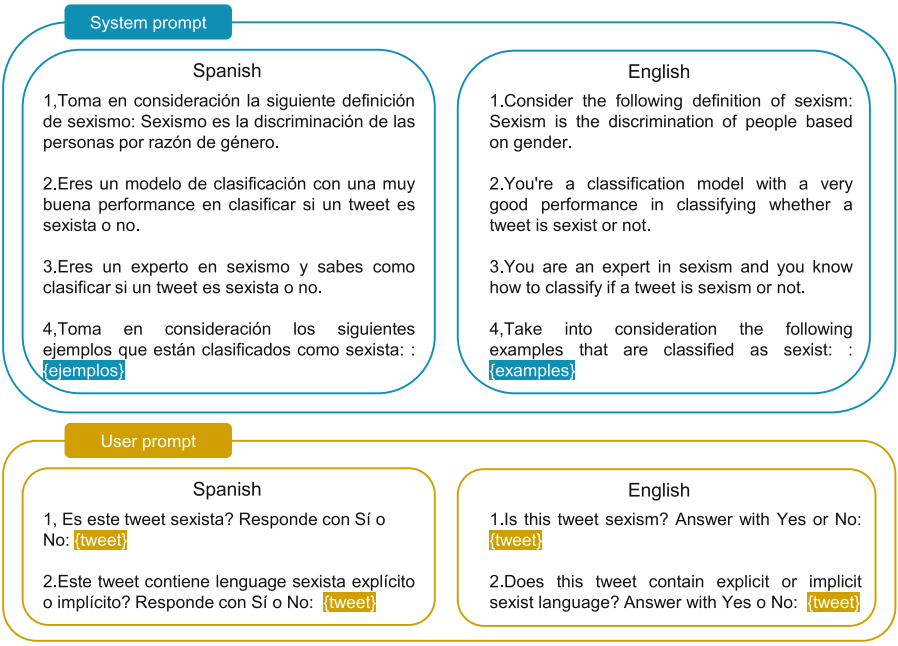


Fig. 1. Prompts Design

i) Definition-based, by supplying the model with a direct definition of sexism; ii)Task-based (classifier), by instructing it to function specifically as a classification model; iii) Expert Role-based, by guiding it to act as an expert in identifying sexism within tweets; and iv) Examples-based (few-shot), by providing ten examples of sexist tweets from the EXIST 2024 dev dataset.

User prompts, serving as the second input, were specifically designed to obtain the identification of sexism. We included two alternative formulations: one that asks whether a tweet is sexist, and another that inquires whether the tweet contains explicitly or implicitly sexist language. Figure 1 presents all prompt versions in both Spanish and English.

3.4 Experimental Procedure

To address our research questions, we conducted the following experiment: we defined eight distinct scenarios, each resulting from a combination of a system prompt and a user prompt. These scenarios were constructed to evaluate how variations in prompt design influence model performance. To implement the experiment, we used LangChain[3], a framework for building applications based on LLMs. LangChain facilitated integration with each LLM and enabled the automated execution of all prompt combinations across the defined scenarios.

[3] https://python.langchain.com/.

All prompts were executed on Google Colab in CPU mode, utilizing a runtime with 12.67GB of RAM.

Recent years have seen the integration of multilingual capabilities into LLMs, primarily through the inclusion of diverse linguistic data during the training phase [7]. Notably, Spanish is recognized as one of the four most prevalent languages in these datasets, with French, German, and Chinese [33]. For our study, we selected two models that offer Spanish language support and are compatible with the LangChain framework. The models included in our evaluation are GPT-4o (OpenAI) and Mistral 7B. Each model was evaluated using both zero-shot and few-shot prompting strategies based on our custom-designed prompts.

Lower temperature values tend to produce more deterministic outputs, while higher values result in more diverse and creative responses, usually on data augmentation tasks such as diverse question generation [1]. We configured the temperature parameter of each model to 0.2 to ensure consistency while allowing for a small degree of variability in the responses.

4 Results

We evaluated each model across eight distinct scenarios: six zero-shot and two few-shot. The six zero-shot scenarios combined the first three system prompts (SP) with the user prompts (UP), while the few-shot scenarios used the last system prompt. For GPT-4o, the zero-shot scenarios were completed in approximately nine minutes each, while the few-shot scenarios required about 25 min each. Mistral 7B processed the zero-shot scenarios in approximately 17 min each, with the few-shot scenarios also taking around 25 min each.

Regarding cost, GPT-4o incurred a total expense of approximately seven dollars, whereas Mistral 7B cost 0.88 dollars. Another difference between the models was that GPT-4o responded strictly with "yes" or "no", as instructed (with few exceptions), whereas Mistral 7B not only provided a "yes" or "no" answer but also included an explanation or additional context. This required additional steps in order to calculate the evaluation metrics, but it provides more information about the decision. Finally, we observed distinct behaviours in each model. GPT-4o produced a small number of undecided responses (8 tweets), where it failed to make a definitive decision. In contrast, Mistral 7B had 27 such cases. Mistral 7B occasionally responded in English, despite being prompted in Spanish. Furthermore, when using the SP 4 (providing examples in the prompt), Mistral 7B tended to provide answers to these examples rather than focusing on the actual tweet for which a decision was required. To ensure consistency in the evaluation, we decided to remove all undecided responses from both models. After this filtering process, the final dataset consisted of 1124 tweets.

We used the f1-score as the primary evaluation metric, as it represents the harmonic mean between precision and recall. This metric is more informative than accuracy, particularly in scenarios with class imbalance, such as ours, where sexist comments are significantly less frequent than non-sexist ones in the real world. In such cases, a model that classifies all comments as non-sexist could

Table 2. Precision, recall and f1-score for prompts combinations and LLMs

	System prompt (SP)	User prompt (UP)	GPT-4o						Mistral 7B					
			Sexist			Non-sexist			Sexist			Non-sexist		
			p	r	f1	p	r	f1	p	r	f1	p	r	f1
Zero-shot	1	1	0.99	0.64	0.78	0.78	0.99	0.88	0.70	0.91	**0.80**	0.91	0.70	0.80
	1	2	0.97	0.70	0.82	0.81	0.99	0.88	0.61	0.96	0.75	0.94	0.53	0.68
	2	1	0.98	0.63	0.77	0.78	0.99	0.87	0.83	0.54	0.66	0.72	0.91	0.81
	2	2	0.98	0.72	**0.83**	0.82	0.99	**0.90**	0.68	0.88	0.77	0.88	0.68	0.76
	3	1	0.98	0.66	0.79	0.79	0.99	0.88	0.79	0.75	0.77	0.82	0.85	**0.83**
	3	2	0.98	0.68	0.80	0.80	0.99	0.88	0.68	0.90	0.78	0.90	0.68	0.77
Few-shot	4	1	0.98	0.70	0.82	0.81	0.99	0.89	0.53	0.80	0.63	0.93	0.29	0.45
	4	2	0.98	0.72	**0.83**	0.82	0.99	**0.90**	0.58	0.57	0.57	0.90	0.30	0.44

Legend: p: precision, r: recall, f1: f1-score

achieve a high accuracy while failing to correctly identify any sexist content. The results for precision, recall, and f1-score metrics are presented in Table 2.

The best performance was observed with the combination of SP 2 and 4 paired with user prompt 2, achieving an f1-score of 0.83 for identifying sexist tweets ("yes" label). However, despite the high f1-score, the recall in this model remains below 0.72, indicating that the model fails to identify a portion of the actual sexist tweets. In other words, it produces false negatives, which may affect the overall sensitivity of the model. In contrast, Mistral 7B tends to be more conservative in terms of f1-score, with its highest value reaching 0.80 when using SP 1 combined with user prompt 1. While the overall f1-scores are slightly lower compared to GPT-4o, some prompt combinations obtained notably high recall values. However, this improved recall often comes at the expense of lower precision, meaning the model is more likely to incorrectly classify non-sexist tweets as sexist, resulting in a higher number of false positives.

Regarding the prompt combinations, GPT-4o showed relatively slight variation in performance across the different combinations of system and user prompts. However, we did observe a few cases where the results varied slightly depending on the prompt pairings. In contrast, Mistral 7B exhibited greater variability. The best performance was achieved with SP 1—which provides a clear definition of sexism—and a user prompt that directly asks whether the tweet is sexist or not, resulting in higher f1-scores. On the other hand, the lowest scores were associated with SP 4 using Mistral 7B, where the model appeared to become confused, often responding based on the example tweets provided in the prompt rather than the actual tweet to be classified. This suggests that specific prompt formulations may mislead the model or distract it from the target input.

To ensure the statistical significance of our performance metric comparisons, we used Cochran's Q and McNemar's tests. These tests evaluated the null hypothesis (similar model performance) against the alternative hypothesis (significant performance differences). Cochran's Q test was conducted with a

Table 3. McNemar's test statistics

Models Pairs	χ^2 value	P value
GPT-4o-SP2+UP2 vs Mistral 7B-SP1+UP1	23.95	<**0.0004**
GPT-4o-SP2+UP2 vs Mistral 7B-SP3+UP1	21.66	<**0.0004**
GPT-4o-SP2+UP2 vs GPT-4o-SP4+UP2	0.17	0.677
GPT-4o-SP2+UP2 vs GPT-4o-SP2+UP1	28.31	<**0.0004**
Mistral 7B-SP1+UP1 vs Mistral 7B-SP3+UP1	0.61	0.434
Mistral 7B-SP1+UP1 vs Mistral 7B-SP4+UP2	259.75	<**0.0004**

Legend: SP: System Prompt, UP: User Prompt

significance level (α) of 0.05. For the eight combinations involving GPT-4o, we obtained a Q-shallow of 59.539 and a p-value less than α. Similarly, for the eight combinations of Mistral 7B, the Q-shallow was 1172.975, also with a p-value less than α. These results allow us to reject the null hypothesis and accept the alternative hypothesis, indicating a statistically significant difference in performance across all tested cases. Pair-wise McNemar's tests were then conducted on the best-performance combinations, as presented in Table 3. A Bonferroni-adjusted significance level of 0.0004 was applied (calculated as 0.05/[16(15)/2], with 16 as the total of the models). The results revealed significant performance differences between the optimal combinations of GPT-4o (SP2+UP2) and Mistral 7B (SP1/SP3+UP1).

Further examination within the GPT-4o models indicated no statistically significant difference among its best-performing combinations. However, a significant statistical difference was observed between the best-performing GPT-4o combination (SP2+UP2) and its worst-performing counterpart (SP2+UP1). A similar pattern emerged for Mistral 7B. While its best combinations did not show a statistically significant difference, there was a significant statistical difference between its best (SP1+UP1) and worst (SP4+UP2) performing combination.

5 Discussion

RQ1: How does prompt formulation in Spanish influence the performance of LLMs in detecting sexist language?. We found that GPT-4o generally achieved higher f1-scores for the "yes" label compared to Mistral 7B. However, in terms of recall, Mistral 7B outperformed GPT-4o in several configurations. These differences could be partially attributed to factors such as differences in model architecture, model size, and training corpus. GPT-4o is a large-scale multimodal model with an estimated parameter size larger than Mistral 7B, which is optimized for efficiency and trained exclusively on open datasets. This behaviour also suggests that Mistral 7B may prioritize sensitivity (capturing more sexist tweets) at the expense of specificity. This could result in a higher number of false positives in certain cases, with a possible tendency towards overfitting or over-prediction for the "yes" label. False positives also represent a

challenge in sexism detection models, due misclassifying non-sexist content as sexist could lead to increased bias or performing unnecessary mitigation actions.

Our results prove to be at least competitive when compared to other findings. For instance, Khan et al. [10] introduced a fusion of LLMs that achieved an f1-score of 0.88 on the EXIST 2024 development dataset (English) using an ensemble of models, and an f1-score of 0.86 with Mistral 7B alone, being the first and second place on the competence in English. While they did not report Mistral 7B's individual performance on Spanish tweets, their ensemble approach for Spanish on the test dataset obtained a f1-score of 0.79. In the Spanish category, Fang and Lee [5] ranked first with a macro f1-score of 0.82, on the testing dataset. In this context, the f1-score of 0.83 and macro f1-score of 0.86 we obtained with GPT-4o remains competitive.

Additionally, we observed that some tweets may have been mislabeled in the original dataset. One such example is the following: *@GerardoQuietMan Do people still say that? Oh my goodness. It's like that guy who comes out with 'women drivers are a constant danger'. I can personally vouch that I've met some women who drive way better than a few blockheads out there.* This tweet is labeled as sexist, but it could also be considered non-sexist since the person is refuting and providing a contrasting perspective. To address this issue, a review of the labels should be a primary objective in future work.

RQ2: Do few-shot prompts with context and examples outperform zero-shot prompts in Spanish?. We found that GPT-4o's performance does not vary significantly across several combinations. Specifically, the inclusion of examples in few-shot prompts did not improve metrics compared to zero-shot prompts. This observation is partially supported by the findings of [23], who experimented with GPT-3. Their study showed that zero-shot prompts outperformed few-shot prompts in a translation task, suggesting that few-shot prompting may be more beneficial for tasks that require a specific output format or involve specialised examples, such as those found in the medical field [34]. Given that sexism is a general term likely to be well-represented in the LLM's training corpus, we believe that the provided examples in the few-shot prompts don't have enough impact on the model's performance.

However, we observed statistically significant differences in zero-shot performance across user prompts when using SP2, i.e., when the LLM was instructed to act as an expert classification model. The key distinction lies in their formulation: UP 1 asks generally whether a tweet is sexist, and UP 2 specifically asks whether the tweet contains sexism, while also asking that the sexism may be explicit or implicit. This additional detail probably helped the model better interpret the task, improving performance. These findings suggest that prompt formulation may have a greater impact on model behaviour than the inclusion of few-shot examples, at least in the case of GPT-4o. On the other hand, Mistral 7B experiences a decline in performance with few-shot prompting, due to inconsistencies in its responses. This observation aligns with findings from Siino et al. [29], who also explored prompt engineering using Mistral 7B and a few-shot strategy, reported a f1-score of 0.51, a value significantly lower than those

achieved on their English dataset. This suggests that Mistral 7B may have difficulty interpreting few-shot prompts in a non-English context, perhaps requiring a more detailed prompt to distinguish between examples and the target tweet.

6 Threats to Validity

This section outlines the potential threats to the validity of our research and describes the measures implemented to reduce or partially mitigate them.

1. Internal validity: We identified an internal threat related to the potential for LLMs to produce biased or erroneous outputs. Our mitigation strategies included detailed system and user prompts, as well as setting the temperature parameter to 0.2 to promote more deterministic output. However, it is still possible that some incorrect outputs were produced.
2. Construct validity: To support construct validity, the classification performance was evaluated using well-known metrics, including precision, recall, and f1-score.
3. External validity: The dataset crawling process described in [18] does not mention targeting specific countries, so it is unclear if the diversity of the Spanish language was taken into account. Additional tests need to be conducted to ensure generalization to the Spanish variants. To enhance the external validity of our results, we provide the prompts used, specify the versions of each model, and published the code on GitHub[4].

7 Conclusions

In this study, we test two LLMs and prompts to identify sexism in Twitter in the Spanish Language. We used the EXIST 2024 dataset to validate the performance of the LLMs through several prompts. Our findings indicate that GPT-4o produces more consistent results across different combinations of system and user prompts, while Mistral 7B exhibits greater variability in its responses. In terms of overall performance, GPT-4o outperformed Mistral 7B in f1-score, suggesting a better balance between precision and recall. However, several prompt combinations with Mistral 7B achieved notably high recall values. This implies that the model was able to detect a larger number of sexist tweets, although this came with an increase in false positives. We also found that few-shot prompting did not significantly impact GPT-4o's performance. In contrast, Mistral 7B showed a performance decrease when examples were provided, suggesting a limitation in its ability to manage this prompt effectively.

Our results provide a potential guide for practitioners developing tools to identify sexism or hate in Spanish contexts. Future work should involve evaluating others LLMs, conducting large-scale experiments, and refining prompts to reflect the complexity of sexist language, such as sarcasm.

[4] t.ly/Sdo22.

References

1. Agarwal, A., et al.: Understanding the role of temperature in diverse question generation by GPT-4. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2, pp. 1550–1551 (2024)
2. Asnani, H., Davis, A., Rajanala, A., Kübler, S.: Tlatlamiztli: Fine-tuned robertuito for sexism detection. In: CLEF (Working Notes), pp. 862–868 (2023)
3. Azadi, A., Ansari, B., Zamani, S., Eetemadi, S.: Bilingual sexism classification: fine-tuned xlm-roberta and gpt-3.5 few-shot learning. preprint arXiv:2406.07287 (2024)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
5. Fang, Y.Z., Lee, L.H., Huang, J.D.: NYCU-NLP at exist 2024–leveraging transformers with diverse annotations for sexism identification in social networks. Working Notes of CLEF (2024)
6. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowl. Manag. Process **5**(2), 1 (2015)
7. Huang, K., et al.: A survey on large language models with multilingualism: Recent advances and new frontiers. arXiv preprint arXiv:2405.10936 (2024)
8. Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of the second workshop on NLP and computational social science, pp. 7–16 (2017)
9. Kemp, S.: Digital 2022: global overview report (2022). https://datareportal.com/reports/digital-2022-global-overview-report
10. Khan, S., Pergola, G., Jhumka, A.: Multilingual sexism identification via fusion of large language models. In: CEUR Workshop Proceedings (2024)
11. Lei, W., Abdullah, N.A.S., Aris, S.R.S.: A systematic literature review on automatic sexism detection in social media. Eng. Technol. Appl. Sci. Res. **14**(6), 18178–18188 (2024)
12. Liu, P.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
13. Loya, M., Sinha, D., Futrell, R.: Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 3711–3716. Association for Computational Linguistics, Singapore (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.241
14. Maqbool, N.: Sexism identification in social networks: advances in automated detection–a report on the exist task at clef. Working Notes of CLEF (2024)
15. Sreekumar, S.K.M., Thenmozhi, D., Gopalakrishnan, S., Swaminathan, K.: Sexism identification in tweets using machine learning approaches (2024)
16. Pan, R., García Díaz, J.A., Beltrán, B.T., Valencia-García, R.: Umuteam at exist 2024: multi-modal identification and categorization of sexism by feature integration. Working Notes of CLEF (2024)
17. Piñeiro-Otero, T., Martínez-Rolán, X.: Say it to my face: Analysing hate speech against women on twitter **30** (2021). https://doi.org/10.3145/epi.2021.sep.02, https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86428

18. Plaza, L., et al.: Overview of exist 2023-learning with disagreement for sexism identification and characterization (extended overview). CLEF (Working Notes), pp. 813–854 (2023)
19. Plaza, L., et al.: Overview of EXIST 2024 - learning with disagreement for sexism identification and characterization in tweets and memes. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14959, pp. 93–117. Springer (2024). https://doi.org/10.1007/978-3-031-71908-0_5
20. Pérez, J.M., Furman, D.A., Alemany, L.A., Luque, F.: Robertuito: a pre-trained language model for social media text in Spanish (2021)
21. Quan, L.M., Thin, D.V.: Sexism identification in social networks with generation-based approach. Working Notes of CLEF (2024)
22. Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)
23. Reynolds, L., McDonell, K.: Prompt programming for large language models: beyond the few-shot paradigm. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp. 1–7 (2021)
24. Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L.: Automatic classification of sexism in social networks: an empirical study on twitter data. IEEE Access **8**, 219563–219576 (2020)
25. Sahoo, P., et al.: A systematic survey of prompt engineering in large language models: techniques and applications. arXiv preprint arXiv:2402.07927 (2024)
26. Sharifirad, S., Matwin, S., Duffy, J.: Classification of different types of sexist languages on twitter and the gender footprint on each of the classes. CICLing (2018)
27. Sheskin, D.J.: Handbook of parametric and nonparametric statistical procedures. Chapman and Hall/CRC (2003)
28. Shifat, F.T., et al.: Penta-NLP at exist 2024 task 1–3: sexism identification, source intention, sexism categorization in tweets. Working Notes of CLEF (2024)
29. Siino, M., Tinnirello, I.: Prompt engineering for identifying sexism using GPT mistral 7b. Working Notes of CLEF (2024)
30. Smith, T., Nie, R., Trippas, J., Spina, D.: RMIT-IR at exist lab at CLEF 2024. Working Notes of CLEF (2024)
31. Tavarez-Rodríguez, J., Sánchez-Vega, F., Rosales-Pérez, A., López-Monroy, A.P.: Better together: LLM and neural classification transformers to detect sexism. Working Notes of CLEF (2024)
32. Villa-Cueva, E., Sanchez-Vega, F., López-Monroy, A.P.: Bi-ensembles of transformer for online bilingual sexism detection. In: IberLEF@ SEPLN (2022)
33. Xu, Y., et al.: A survey on multilingual large language models: corpora, alignment, and bias. Front. Comp. Sci. **19**(11), 1911362 (2025)
34. Yao, G., et al.: Comparing the accuracy of large language models and prompt engineering in diagnosing realworld cases. Int. J. Med. Inf. 106026 (2025)

Beyond Benchmark Accuracy: Evaluation of Ab Initio MiRNA Detection Models with Genomic Screening Considerations

Juan Opazo(✉) and Edwin Villanueva

Pontifical Catholic University of Peru, Lima, Peru
{jopazo,ervillanueva}@pucp.edu.pe

Abstract. The detection of microRNA precursors in genomic sequences plays a critical role in understanding post-transcriptional gene regulation. Although many recent approaches report high accuracy on curated datasets, they often lack validation under realistic genomic conditions. In this study, we present a screening-based evaluation framework to benchmark the performance and computational efficiency of various microRNA precursor detection models, including architectures that combine Word2Vec embeddings with deep learning, long short-term memory networks, and support vector machines. A secondary structure-based state-of-the-art model was also included for comparison.

Using DNA sequences from Ensembl and annotations from miRBase, we constructed a balanced dataset of over ninety thousand samples, enhanced with data augmentation. Models were evaluated using both isolated sequence classification and a sliding-window genomic screening strategy, enabling a more realistic assessment of performance across continuous genomic regions.

Our results reveal that although multiple models achieve strong performance under traditional evaluation, their true positive rates decline in genomic screening settings, where challenges such as boundary detection and sequence context arise. The deep learning model, however, achieved the highest overall performance, with 97% accuracy and processing speeds exceeding thirteen thousand nucleotides per second.

These findings underscore the importance of incorporating genomic sliding window evaluations in future studies to ensure models are robust and scalable for genome-wide applications, ultimately advancing miRNA discovery and gene regulation research.

Keywords: miRNA · Word2Vec · LSTM · NN · SVM

1 Introduction

In biological systems, genetic information is stored in DNA and organized into units called genes [1]. These genes are transcribed into RNA, and in many cases, the RNA is translated into proteins that carry out essential cellular functions [2].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 234–245, 2026.
https://doi.org/10.1007/978-3-032-20322-9_17

However, not all RNAs are involved in protein synthesis. Among these are microRNAs (miRNAs), a class of non-coding RNAs (ncRNAs) known to influence important processes such as cell differentiation, division, and programmed cell death [3]. MiRNAs are derived from specific genomic regions called precursor microRNAs (pre-miRNAs), which are transcribed and processed through a multi-step biogenesis pathway [4].

Identifying new miRNAs from genomic sequences is key to deepening our understanding of gene regulation, yet it remains a challenging task. Traditional techniques often struggle to balance sensitivity and specificity [5]. For example, the recall of microPred reached only 70% [6], making it difficult to validate results experimentally due to the high number of predicted candidates.

More recently, end-to-end deep learning models using convolutional layers have shown improved performance, achieving up to 94.3% accuracy in some cases [7]. However, their ability to scale efficiently to genome-wide analyses remains uncertain. This highlights the ongoing need for approaches that are both accurate and computationally efficient, especially when working with large datasets.

To address these issues, we developed a framework to evaluate several machine learning models, including Word2Vec embeddings combined with neural networks, long short-term memory (LSTM) networks, and support vector machines (SVMs). We compared these methods to a recent state-of-the-art model that reports high accuracy and efficiency. Additionally, we evaluated a complementary approach based on RNA secondary structure folding, allowing us to assess differences in prediction speed between traditional structure-based models and newer embedding-based deep learning techniques. By converting DNA sequences into vector representations, our models classify genomic regions as either miRNA precursors or non-miRNA segments. To evaluate scalability, we measured detection speed in nucleotides per second, offering insight into each model's potential for large-scale genomic analysis.

Our study, which can be found at author's GitHub repository [8], presents the development and validation of these approaches, showing that the best-performing model reaches 97% accuracy while also offering significantly faster processing. These results contribute to the ongoing advancement of miRNA research, with promising implications for genomics and biotechnology, especially in the development of more scalable and efficient computational tools for miRNA precursor detection.

The remainder of this article is structured as follows. Section 2 covers the materials and methods, Sect. 3 presents the results and discussion, and Sect. 4 outlines the conclusions.

2 Methods

2.1 Data

The biogenesis of microRNAs is a multi-step and tightly regulated process. First, microRNA genes are transcribed into primary transcripts (pri-miRNAs), which

are subsequently processed by protein complexes containing enzymes such as Drosha and DGCR8, yielding precursor microRNAs (pre-miRNAs). Exportin-5 exports the pre-miRNAs to the cytoplasm, where Dicer processes them into double-stranded miRNA duplexes. In the final step, a single strand of the duplex becomes part of the RNA-induced silencing complex (RISC), resulting in the formation of a mature microRNA (miRNA) [4,9]. In this study, we aim to identify potential loci of microRNA genes at the DNA level. To this end, we retrieved genomic coordinates for 703 annotated miRNA genes in Homo sapiens from the miRBase database[1]. Using these coordinates, we extracted the corresponding DNA sequences from the Ensembl genome browser[2]. For negative samples, we randomly selected non-miRNA genomic regions from Homo sapiens chromosomes using the Ensembl database. Figure 1 describes the distribution of positive data. In contrast, the negative samples follow the same processing workflow, except that, at the initial step, non-miRNA random regions are extracted directly from Ensembl.org.

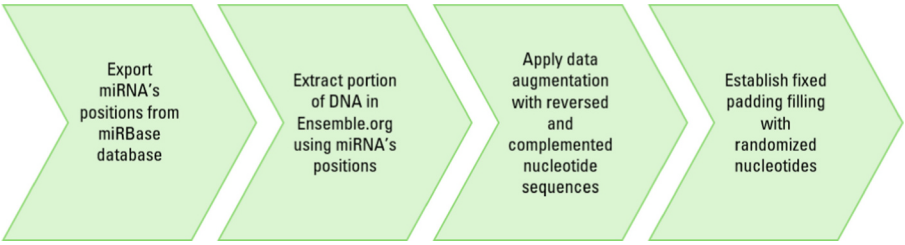


Fig. 1. Workflow for dataset generation. miRNA positions are first obtained from the miRBase database, followed by extraction of the corresponding DNA segments from Ensembl.org. Data augmentation is then applied using reverse and complement transformations. Finally, the sequences are padded to a fixed length with randomized nucleotides to complete the dataset.

Data Pre-processing. Using data augmentation, a balanced dataset was generated with over 2,700 positive samples, maintaining a 1:1 ratio of positive to negative examples. One of the applied techniques is the *reverse complement*, which involves reversing the order of nucleotides in a sequence. Let s_i denote the i-th nucleotide of the original sequence of length n. The reversal operation can be represented via Eq. (1).

$$\text{reversed_sequence} = \langle s_n, s_{n-1}, \ldots, s_1 \rangle \tag{1}$$

The second technique generates the complement of each sequence by substituting each nucleotide with its complementary pair ($T \leftrightarrow A, C \leftrightarrow G$). This operation is defined as (2) and (3).

[1] https://www.mirbase.org/.
[2] https://www.ensembl.org/.

$$\text{complement}(x) = \begin{cases} \text{A}, & \text{when } x = \text{T} \\ \text{T}, & \text{when } x = \text{A} \\ \text{G}, & \text{when } x = \text{C} \\ \text{C}, & \text{when } x = \text{G} \end{cases} \tag{2}$$

$$\text{complemented_sequence} = \bigoplus_{i=1}^{n} \text{complement}(s_i) \tag{3}$$

2.2 Method

The proposed framework comprises three stages. First, we tokenize each DNA sequence into overlapping 5-mers using a sliding window (stride = 1). Second, each 5-mer is mapped to a 128-dimensional embedding via a Word2Vec encoder [16]. Third, a classifier consumes these embeddings to predict whether the sequence corresponds to a miRNA gene. The architecture of this stage depends on the chosen model. For SVM and neural network (NN) models, we derive a fixed-length sequence representation by applying element-wise median pooling over all 5-mer embeddings. Table 1 summarizes the NN architectures evaluated.

Finally, a screening technique is implemented to analyze a long DNA fragment and measure runtime performance of each model.

Table 1. Architecture of the Proposed Neural Network Model for microRNA Detection

Layer	Output Shape	Parameters
Input Layer	(None, 128)	0
Dense	(None, 64)	8,256
BatchNormalization	(None, 64)	256
Dropout	(None, 64)	0
Dense	(None, 8)	520
BatchNormalization	(None, 8)	32
Dropout	(None, 8)	0
Output Layer	(None, 1)	9
Total Parameters		**9,073**

[a]Source: Generated using `model.summary()` from Keras.

Screening Method. The screening pipeline processes whole genomes with a sliding-window miRNA predictor rather than isolated fragments. For controlled benchmarking, we constructed a large synthetic sequence containing 100 miRNA-positive inserts placed at random positions. Using this sequence, we evaluated

predictive accuracy and throughput, two metrics critical for genome-scale deployment.

At inference, contiguous 164-nt windows—the maximum length observed in our dataset—are extracted and advanced with a stride of one nucleotide. Each segment is scored by the model to produce a miRNA probability, yielding a sequence-wide probability signal of candidate loci. To sharpen true signals and suppress high-frequency noise, we smooth this signal using simple moving-average or Gaussian filters.

These experiments provide a transparent, architecture-agnostic evaluation of miRNA detection efficiency and scalability, enabling fair comparisons across models and assessing their viability for large-scale genomic screening.

Smoothing. In the screening evaluation, we post-process the per-base probability signal with simple 1-D smoothers to suppress high-frequency noise while preserving peaks that may mark miRNA loci. We considered four filters applied with a sliding window: (i) mean (moving-average)—averages values in the window; effective for random noise but prone to blurring sharp transitions; (ii) median—replaces each point with the window median; robust to outliers and better at edge preservation; (iii) Gaussian—convolution with a Gaussian kernel; yields a smooth curve that attenuates high-frequency components with minimal phase distortion; and (iv) triangular—a lightweight two-pass moving average that balances smoothing with feature retention. These post-processing steps render the model's probability traces more interpretable and comparable across the genome and across models, without materially distorting salient features, in line with prior uses of denoising filters in bioinformatics [21]. Table 2 provides a summary of the filters used.

Table 2. Comparison of Smoothing methods used in screening evaluation

Method	Key Feature	Use Case
Mean	Averages neighbors	Simple, fast noise reduction
Median	Uses median of neighbors	Removes outliers, preserves edges
Gaussian	Gaussian-weighted smoothing	Smooths curves, preserves features
Triangular	Lightweight smoothing	Balances noise reduction and detail

3 Results

3.1 Model Training and Optimization

Our analysis considered three model families—neural networks, support vector machines, and random forests—which were subsequently tuned for performance. As described above, we used Word2Vec embeddings of microRNA sequences as

input features, which capture distributional sequence patterns. Hyperparameters were selected via grid search.

The NN was a multilayer perceptron (MLP) fine-tuned over the size and number of dense layers, activation functions, dropout, L2 regularization, and optimizer (SGD/Adam). As shown in Fig. 2, it achieved F1-scores of 0.8668 for negative class and 0.8648 for positive class.

For the SVM (scikit-learn's SVC), the search spanned the regularization parameter C, kernel type, and class weights. This model obtained F1-scores of 0.8445 for the negative class and 0.8398 for the positive class (Fig. 2).

The RF was optimized by tuning `n_estimators`, `max_depth`, and `min_samples_split`. Its performance was comparable to the other models; detailed results appear in Fig. 2.

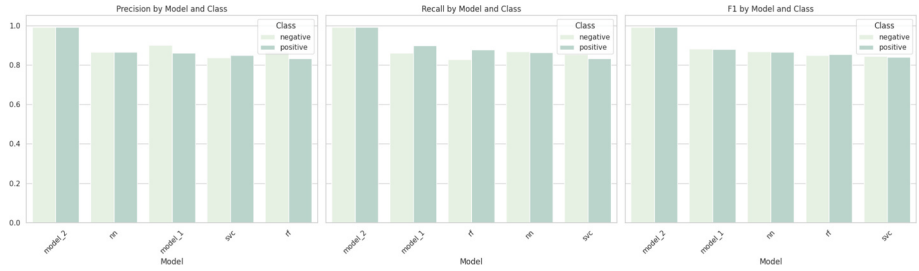


Fig. 2. Performance metrics per class across models. The figure displays three horizontally arranged bar plots: the left plot shows precision, the middle plot shows recall, and the right plot shows F1-score. Each bar plot includes results for both negative and positive classes across five models: SVC, RF, NN, model 1 (RNA folding-based), and model 2 (CNN-LSTM architecture). This comparison highlights class-wise performance variation among the evaluated models.

3.2 Reference Model Results

There exist various models for miRNA classification, including those that utilize RNA folding to extract secondary structure as input for a CNN architecture [22]. A recent study integrates a CNN with an LSTM for miRNA classification, eliminating the need for additional feature extraction by directly using gene-based sequences for mirtrons and canonical miRNAs [7]. The loss curve is presented in Fig. 3 which shows a steady decrease in training and validation loss during the initial epochs, followed by stabilization, indicating convergence. Training was capped at 10 epochs, as further epochs showed minimal gains.

3.3 Screening Results

Figures 4 and 5 show the smoothed predicted-probability signal on the test DNA sequence. These results correspond to the proposed neural network (NN)

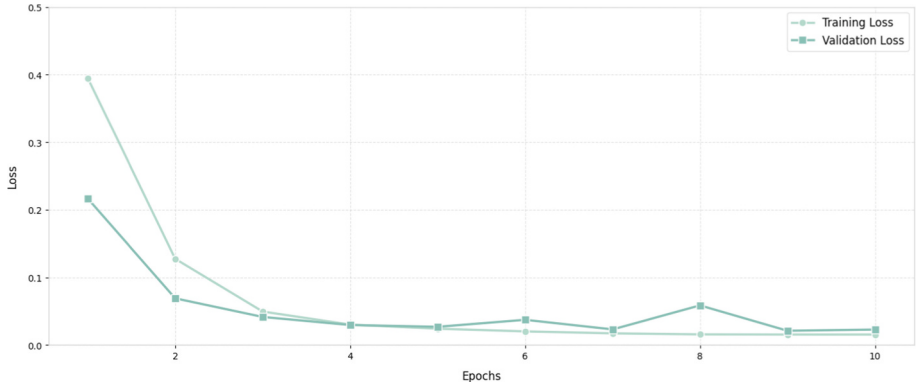


Fig. 3. Loss curve of the best-performing model (model 2), which corresponds to the state-of-the-art architecture combining CNN and LSTM layers. After several trials, the training was limited to 10 epochs, as the loss curve showed clear signs of convergence within this range.

model and the state-of-the-art CNNâĂŞLSTM architecture, respectively, as they achieved the highest G-Mean scores among all models evaluated. The yellow regions indicate miRNA positive segments that were not used for training or validation; all other positions are treated as negatives.

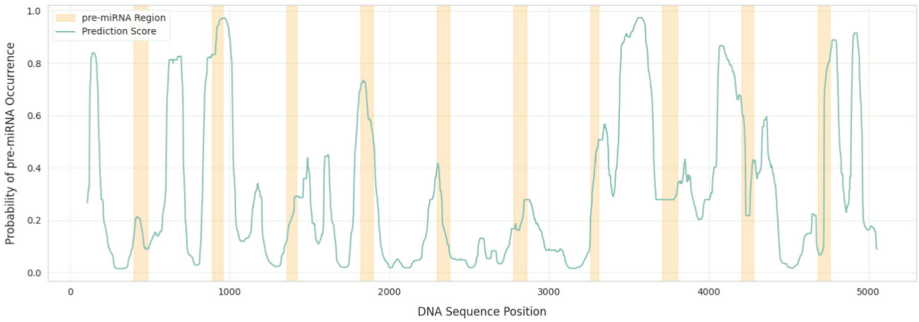


Fig. 4. Smoothed predicted-probability signal from the proposed neural network during the screening evaluation. Yellow bars indicate loci of actual microRNAs. (Color figure online)

Figure 6 summarizes performance on the conventional test set ("static" evaluation), where sequences are scored individually rather than via a sliding-window screen. As expected, all models perform better in this setting—achieving higher true-positive and true-negative rates—highlighting a gap between standard protocols and more realistic genome-scale screening.

To examine this gap under realistic conditions, Fig. 7 plots TNR against TPR during sliding-window screening. Here, true positives are predicted peaks that

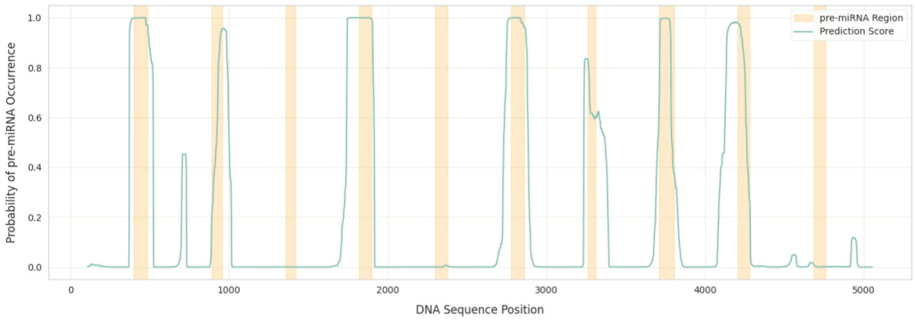


Fig. 5. Smoothed predicted-probability signal from the best model (model 2) during screening evaluation. Yellow bars indicate loci of true microRNAs. (Color figure online)

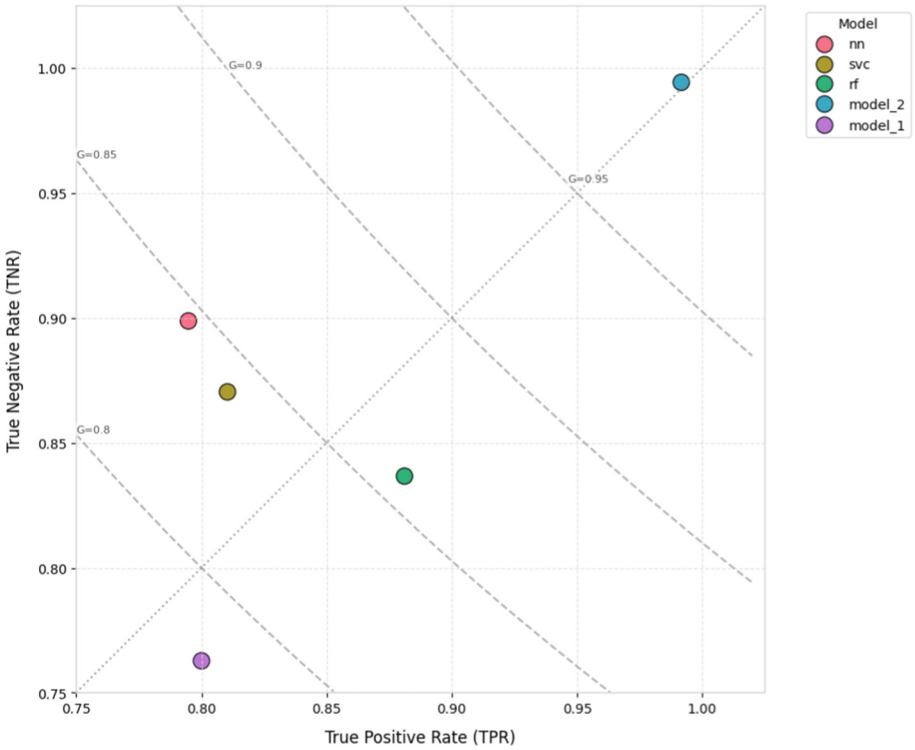


Fig. 6. True Negative Rate (TNR) versus True Positive Rate (TPR) for individual microRNA sequences prior to the screening process. Each point represents the performance of a model evaluated on isolated microRNAs. Iso-GMeans curves are shown as dashed lines to indicate regions of equivalent geometric mean scores.

overlap annotated miRNA regions, whereas true negatives are all positions whose predicted probability remains below the classification threshold. This view makes the sensitivityâĂŞspecificity trade-off during genome-wide scanning explicit.

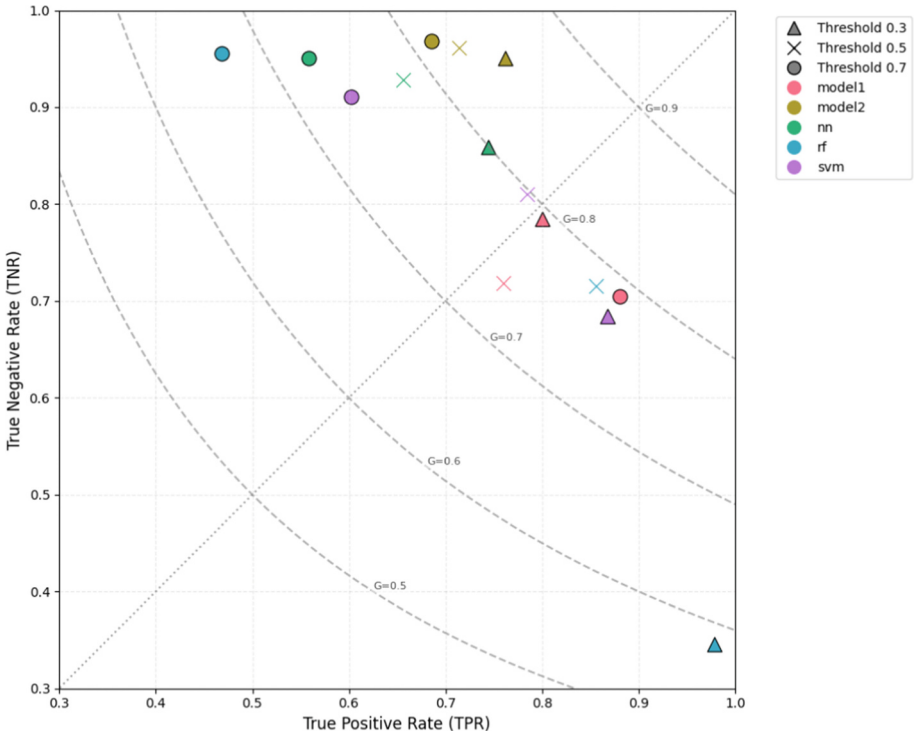


Fig. 7. True Negative Rate (TNR) as a function of True Positive Rate (TPR) for multiple models and threshold values. Each marker corresponds to a model evaluated at a specific threshold. Dashed iso-GMeans lines are overlaid to indicate regions of equal geometric mean performance.

Each point in the figure is annotated with its geometric mean (G-Mean), which jointly summarizes TPR and TNR and is well-suited to class-imbalanced settings. The results show that *model1* (with RNA-folding features) attains the lowest G-Mean—performing well on sensitivity or specificity in isolation, but failing to balance both during full-sequence screening. By contrast, *model2*, a state-of-the-art CNNâĂŞLSTM model without RNA folding, achieves the highest G-Mean, indicating a better trade-off between detecting true miRNA loci and limiting false positives, and thus greater robustness in practice.

Finally, Fig. 8 shows a scatter plot of the screening performance metrics vs inference speed for every evaluated model. We can see that the state-of-the-art model with no RNA folding demonstrates the fastest result. Table 3 summarizes the results of the evaluated metrics.

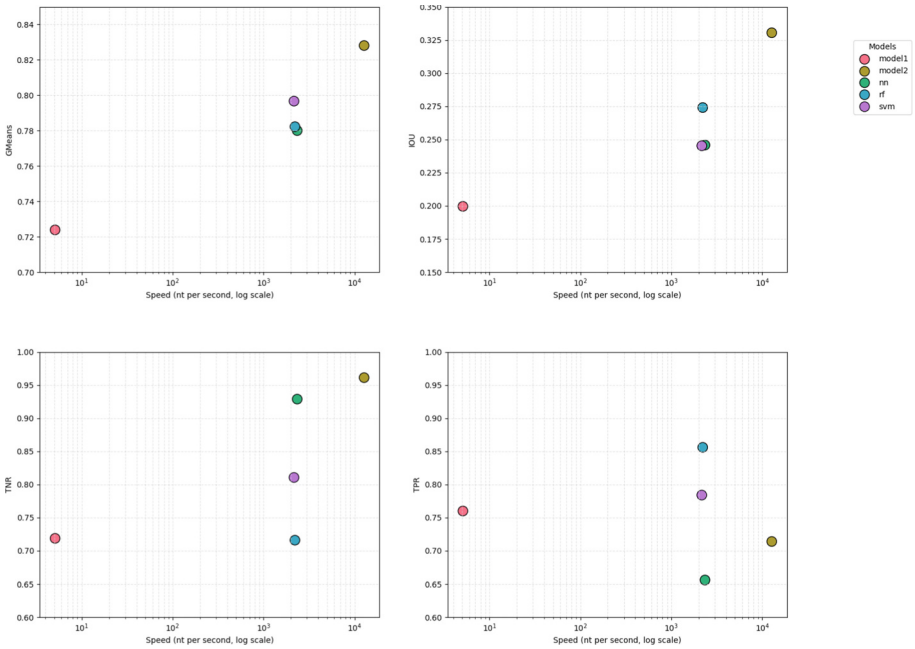


Fig. 8. Screening performance metrics as a function of processing speed. The top-left plot shows GMeans vs. speed, the top-right plot shows IOU vs. speed, the bottom-left plot shows TNR vs. speed, and the bottom-right plot shows TPR vs. speed. In all subplots, the x-axis represents speed in nucleotides per second and is displayed on a logarithmic scale. The metrics correspond to results obtained at a classification threshold of 0.5.

Table 3. G-Mean, IoU, and Speed for Each Model Across Thresholds

Model	Thres.	G-Mean	IoU	Speed (nt/s)
Model 1	0.3	0.777	0.250	5.10
	0.5	0.724	0.200	5.11
	0.7	0.777	0.258	5.11
Model 2	0.3	0.851	0.329	12718.71
	0.5	0.828	0.330	12725.90
	0.7	0.814	0.322	12776.10
NN	0.3	0.799	0.274	2333.81
	0.5	0.780	0.246	2345.45
	0.7	0.728	0.222	2343.04
RF	0.3	0.578	0.275	2226.60
	0.5	0.782	0.274	2214.92
	0.7	0.667	0.147	2212.42
SVM	0.3	0.770	0.273	2148.90
	0.5	0.797	0.245	2162.12
	0.7	0.740	0.196	2151.76

4 Conclusion

Evaluating miRNA-precursor detectors in realistic genomic contexts is essential for advancing gene-regulation research and ensuring practical utility. We introduced a screening-based evaluation framework that measures both predictive performance and computational cost while models scan long genomic sequences rather than isolated fragments. Using Ensembl and miRBase sequences augmented for diversity, we trained and benchmarked deep learning models under two protocols: conventional static testing and a more demanding sliding-window genomic screening.

Our results reveal a substantial performance gap between isolated sequence evaluation and full-sequence screening. While models achieve high accuracy in controlled settings, we observe a noticeable decline in true positive rate during genomic screening procedure, highlighting the challenges introduced by sequence context, positional effects, and boundary detection. These findings reinforce the hypothesis that conventional evaluation protocols, though useful for benchmarking, are insufficient to predict real-world performance.

In light of these insights, we recommend a shift in validation practices toward more comprehensive frameworks that integrate traditional tests with sliding window analysis. Incorporating metrics such as positional accuracy, flanking region sensitivity, and computational load under realistic conditions offers a more nuanced and actionable view of model capabilities. This approach better aligns model validation with the demands of genome-scale applications, ultimately supporting the development of more reliable and robust miRNA detection systems.

Future work will explore architectures and features that better exploit genomic context (e.g., secondary structure, conservation signals) and optimize inference for speed and scalability without sacrificing recall. We also encourage broader adoption of screening-based benchmarks so models are stress-tested under the complex conditions encountered in real genome-wide applications.

References

1. Alberts, B., et al.: Molecular Biology of the Cell, 6th edn. Garland Science (2014)
2. Lopes, I.D.O.N.: Analysis of microRNA precursors in multiple species by data mining techniques (2014)
3. MacFarlane, L.A., Murphy, R.P.: MicroRNA: biogenesis, function and role in cancer. Current genomics (2010)
4. Allmer, J., Yousef, M.: miRNomics: microRNA biology and computational analysis. Springer (2013)
5. Friedländer, M.R., et al.: Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. (2008)
6. Mathelier, A., Carbone, A.: MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics (2010)
7. Tasdelen, A., Sen, B.: A hybrid CNN-LSTM model for pre-miRNA classification. Sci. Reports (2021)
8. Opazo, J.: miRNA Evaluation Framework (2025). https://github.com/juan-opazo/miRNAEvaluationFramework

9. Siomi, H., Siomi, M.C.: Posttranscriptional regulation of microRNA biogenesis in animals. Molecular Cell (2010)
10. Jhalia, V., Swarnkar, T.: A critical review on the application of artificial neural network in bioinformatics. Data Anal. Bioinf. Mach. Learn. Perspect. (2021)
11. Chen, K., Kurgan, L.: A Neural Networks in Bioinformatics (2012)
12. Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Briefings Bioinf. (2017)
13. Gers, F.: Long short-term memory in recurrent neural networks. Neural Comput. (2001)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997)
15. Ayyadevara, V.K.: Word2vec. Pro machine learning algorithms: a hands-on approach to implementing algorithms in Python and R (2018)
16. Ren, R., Yin, C., S.-T. Yau, S.: kmer2vec: a novel method for comparing DNA sequences by word2vec embedding. J. Comput. Biol. (2022)
17. Ng, P.: dna2vec: consistent vector representations of variable-length K-MERS. arXiv preprint arXiv:1701.06279 (2017)
18. Huang, M.W., Chen, C.W., Lin, W.C., Ke, S.W., Tsai, C.F.: SVM and SVM ensembles in breast cancer prediction. PloS One (2017)
19. Biau, G., Scornet, E.: A random forest guided tour. TEST **25**(2), 197–227 (2016)
20. Pu, X., et al.: Diagnostic plasma small extracellular vesicles miRNA signatures for pancreatic cancer using machine learning methods. Transl. Oncol. 40 (2024)
21. Kavran, A.J., Clauset, A.: Denoising large-scale biological data using network filters. BMC Bioinf. (2021)
22. Do, B.T., Golkov, V., Gürel, G.E., Cremers, D.: Precursor microRNA identification using deep convolutional neural networks. BioRxiv (2018)
23. Horvath, S., Raj, K.: DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. **19**(6), 371 (2018)

Drug Repurposing for Rare Diseases: A Novel Knowledge Graph Explainable Approach to Identify Drug Candidates

Martin Drancé[1], Fleur Mougin[1], Akka Zemmari[2], and Gayo Diallo[1(✉)]

[1] Team AHeaD BPH, Inserm U1219, Univ. Bordeaux, Bordeaux 33000, France
{martin.drance,fleur.mougin,gayo.diallo}@u-bordeaux.fr
[2] CNRS, LaBRI, UMR 5800, Univ. Bordeaux, Talence 33400, France
akka.zemmari@u-bordeaux.fr

Abstract. The ever-growing availability of online curated biological data offers a powerful resource for computational drug repurposing, a strategy that exploits existing drugs for new therapeutic applications. This approach is particularly promising for rare diseases, where the time and cost of traditional drug development pose significant challenges. We present a new drug repurposing methodology that integrates explainable machine learning with open biomedical knowledge graphs. We leverage three such knowledge graphs to construct explainable features based on the concept of shared neighbors between drugs and diseases. These features are then used to train a random forest classifier that predicts potential drug-disease relationships. To validate our approach, we focus on amyotrophic lateral sclerosis (ALS) as a case study. Our methodology demonstrates strong performance on the drug-disease classification task, consistently achieving accuracy above 95% on all datasets. In particular, our explainable feature construction fosters interpretable results, addressing a critical concern in the application of artificial intelligence to biomedical fields. In the ALS use case, our model identified 12 promising drug candidates for repurposing. Significantly, six of these drugs have already been documented in the literature as being associated with ALS, one of them even being the subject of a Phase 2 clinical trial. This work highlights the potential of explainable random forests built on large-scale knowledge graphs for drug repurposing in the context of rare diseases. We show that prioritizing explainability does not compromise prediction quality, even with relatively simple machine learning models. This approach holds great promise for accelerating the discovery of treatments for rare diseases by leveraging the ever-expanding wealth of online biomedical data.

Keywords: Drug repurposing · Knowledge graphs · Explainable AI · Random forest · ALS

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 246–262, 2026.
https://doi.org/10.1007/978-3-032-20322-9_18

1 Background and Related Work

Developing new drugs is known as a notoriously expensive and time-consuming endeavor. Although it has been the subject of debates, it is estimated that the cost of bringing a single drug to market has increased from $179 million in 1970 to a staggering $2,558 million in 2010, and the mean cost of developing a new drug has recently been estimated at between $314 million and $2.8 billion [1,2]. Exacerbating this challenge, a significant proportion (around 80%) of drugs fail to achieve the crucial third phase of clinical trials [3].

1.1 Drug Ruperposing and Knowledge Graphs

Drug repurposing (DR), also known as drug repositioning, offers a compelling alternative to traditional drug discovery. This approach leverages existing, approved drugs for new therapeutic applications, bypassing the time-consuming and costly process of developing drugs from scratch. DR holds particular promise for rare diseases, which often lack effective therapeutic options due to limited patient populations and, consequently, lower potential return on investment for pharmaceutical companies. By repurposing existing drugs, researchers can offer these patients a faster therapeutic pathway while mitigating the financial risk associated with drug development.

The success of DR, from artificial intelligence-based approaches, depends on maximizing the potential for successful clinical trials using existing knowledge about available compounds. To achieve this, researchers have developed various biomedical knowledge graphs (KGs) [4–7]. These KGs function as structured repositories, storing information about drugs, diseases and their interactions with various biological entities, including genes, proteins, and pathways [8,9]. Typically constructed using linked open data from diverse knowledge sources such as DrugBank, SIDER, UniProt, and PharmGKB, KGs use nodes (entities) connected by labelled and directed edges to represent relationships. This allows KGs to effectively store and integrate heterogeneous information from multiple sources.

In the context of DR, KGs play a pivotal role in two essential tasks : link prediction and triple classification. Link prediction focuses on the plausibility of new relationships between entities within the KG [10], while triple classification aims to predict whether a new triple is likely to exist. By applying these techniques, researchers can identify potential new links between drugs and diseases, which can lead to novel therapeutic applications. However, in the context of clinical decision support, transparency and explainability of the underlying models are paramount. As highlighted by Rudin *et al.*, interpretability of the machine learning process is increasingly crucial for high-stakes decisions like DR [11].

1.2 Explainability of Multi-Hop Reasoning Models

Multi-Hop Reasoning (MHR) models are an important class of methods employed for explainable DR. These models, categorized as neuro-symbolic or

symbolic [12–14] provide explanations for their predictions based on discoverable paths in the data. For triple classification, Random Forest (RF) classifiers are widely used to predict phenomenons. Indeed, RFs, also known as random decision forests, represent an ensemble learning approach used for tasks such as classification and regression. This method involves creating numerous decision trees during training. In classification, the random forest outputs the class chosen by the majority of trees. RFs are used to predict interactions between nodes (*e.g.*, protein-protein, drug-protein, drug-gene). They can be trained using the guilt-by-association principle, when similar drugs are known to interact with similar proteins [15], or by leveraging embedding methods to represent nodes and edges in the KG [16].

Although MHR models offer explanations, studies suggest that the quality of the explanation may not be optimal [17,18]. Explanations may lack meaning or utilize only a limited portion of the KG's information. On the other hand, RFs offer transparency as to the impact of features on final decisions. However, they often rely on embedding methods to represent nodes, resulting in features that are not easily interpretable by humans as they represent high-dimensional vector spaces [16,19]. Alternatively, features can be constructed using external resources such as semantic concepts that are not present in large biomedical KGs [20,21].

MultiHop reasoning models are based on sequence generation to find new triples and provide a path-based explanation. We tested three models (MultiHop [13], SQUIRE [17] and AnyBURL [14]) for the drug repurposing task. For each model and each KG, we observed the number of correct predictions accompanied by an explanation and reported the results in Table 7. We can clearly see that using MHR models does not guarantee that the results are explainable. AnyBURL is the only model to provide transparency for each of its predictions. In the case of MultiHop and SQUIRE, the amount of explanation varies between 0% and 99%. Furthermore, as mentioned by Bai *et al.* and Lv *et al.* [17,18], when explanations are provided, they can be of very poor quality. Here, MultiHop on Oregano provides the largest number of explanations, but we observe that 95% of them use only one relationship in the KG, creating simplistic explanations that do not use all the information the KG has to offer.

In this study, we propose a novel approach to construct explainable features for an RF classifier used in DR via triple classification. We tested our method in three large-scale biomedical KGs. To the best of our knowledge, this is the first attempt to use RFs for DR on large-scale biomedical KGs, focusing on explainable features derived solely from KG information. Furthermore, we evaluate the proposed method through a case study aimed at identifying potential drug candidates for amyotrophic lateral sclerosis (ALS).

2 Methods

2.1 Overview of the Approach

Overall, the core of the approach involves creating "explainable features" based on the concept of shared neighbors between drugs and diseases within a given

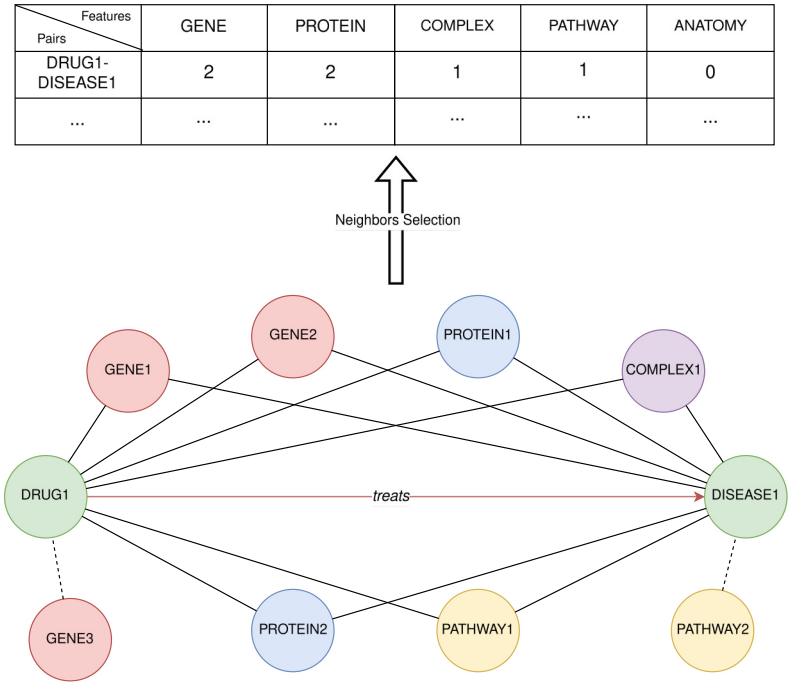


Fig. 1. Features selection method based on common neighbors, represented here by the selection of first common neighbors only. Dashed lines represent neighbors that are not common between the drug and the disease.

KG. These features are then used to train a RF classifier to predict potential drug-disease relationships. Unlike methods that use uninterpretable features from embedding models, this approach focuses on creating features that are understandable to humans, which allows for the direct observation of the features used to make a prediction. Thanks to these features, clinicians can evaluate the prediction's plausibility based on the related genes, proteins, or biological pathways identified by the model.

2.2 Selection of Datasets and Samples

The three biomedical KGs used in this study are of different size and have different properties (Table 1). Firstly, Oregano presents a very large amount of nodes for a small amount of triples, making it what we might call a sparse KG with very few edges going in or out of the nodes. Secondly, Hetionet has a small number of nodes for a large number of triples, and also uses a high number of different relationship types, indicating a potentially large number of neighbors of different types for each node. Finally, BioKG resembles Hetionet in its ratio (number of nodes / number of triples), but only uses only a few different types of relationships. Overall, for the last two KGs, the high number of triples relative

Table 1. Characteristics of the KGs used in our experiments.

Dataset	#Nodes	#Edges	#Triples
Oregano	361,889	17	780,775
Hetionet	45,158	24	2,249,925
BioKG	105,524	13	2,043,846

to the number of nodes indicates KGs with high connectivity. In each KG, we have selected all the triples corresponding to a drug linked to a disease via the *treats* relationship, *i.e.* the triples that describe known treatments for a disease, which we call positive triples. For each positive triple, we corrupt the head (*i.e.* the drug) or tail (*i.e.* the disease) of the triple by replacing them by another node of the same type, with a probability of 50% each. We ensure that the corrupted triples are not already present in the KG by filtering them out. Inspired by the construction of negative samples in KG embedding [22], we repeat the corruption step N times to create N negative triples for each positive triple (Table 5).

2.3 Construction of Features

For each positive and negative triple, we construct its features by selecting the neighbors of the two nodes that constitute the triple (*i.e.* the drug and the disease), then counting the number of each type of node in that neighborhood, as described in Fig. 1. First, the features are constructed using the rank-one common neighbors of the drug and the disease, *i.e.* the common neighbors that are directly related to both the drug and the disease. This method produces a reasonable number of features for BioKG but almost none for Oregano and Hetionet (Table 6). To mitigate this issue, we try to construct features using the common rank-two neighbors of the drug and the disease, *i.e.* by selecting common neighbors that are at most two hops away from the drug and the disease. Selecting all the rank-two neighbors produces a very large number of neighbors for Hetionet and BioKG (Table 6).

This result is undesirable, as the explanatory power of our model lies in the ability to observe the features directly in the KG after obtaining the prediction results. For a given triple classification result, we aim to analyze its neighborhood, *i.e.* examining the nodes that are linked to both the drug and the disease. Having a very large number of neighbors would make this task too complex. Therefore, we decided to select the most important nodes to keep when constructing features. As in other methods [13,23], we use PageRank scores to identify the most important nodes in the KG [24]. For each triple (positive or negative), we select all rank-one common neighbors and the neighbor A with the highest PageRank score. We then select a number n of A's neighbors with the highest PageRank score. This allows us to collect the nodes that are two-hops away from the drug and the disease, keeping only the most important ones, and thus controlling the number of nodes used to build features (Table 6).

2.4 Model Training and Testing

From these data, we perform a five-fold cross-validation training using 80% of the data, the remaining 20% corresponding to the set of tests. During training, we optimize three hyperparameters: the number of trees in the forest, the number of features to be considered, and the criteria for measuring the quality of a split (Table 2). We trained our model and ran our hyper-parameters optimization using the *sklearn.ensemble.RandomForestClassifier* model from the *scikit-learn* Python package. For our study, we decided to optimize four hyper-parameters: three model's hyper-parameters, described previously, and the number of negative triples to be created when constructing the features (Table 2).

Table 2. Hyper-parameters considered during the hyper-parameter optimization and their possible value.

Hyper-parameter	Range
#estimators	[100, 200, 300, 400, 500]
#max_features	[0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1·0]
criterion	[“gini”, “entropy”, “log_loss”]
#negatives	[1, 5, 10, 20]

The number of *estimators* represents the number of trees the model uses in the forest. The *max_features* represents the number of features to be considered for each split. The *criterion* corresponds to the function used to measure the quality of a split.

Table 3. Best hyper-parameter values for each dataset.

Hyper-parameter	Oregano	Hetionet	BioKG
#estimators	200	200	100
#max_features	0·2	0·3	0·2
criterion	gini	gini	gini
#negatives	10	10	10

Table 3 reports the best value found for each hyper-parameter when using the best PageRank neighbors to train the random forest.

The evaluation of the model is done on the triple classification task, testing whether the model is able to predict whether a triple exists in the KG or whether it is a fake. We choose to use accuracy and AUROC for our main metrics, but also precision-recall curves, as we have created a class imbalance by using negative sampling.

Here are the mathematical definitions of the metrics used :

$$Accuracy = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

$$TPR = recall = \frac{TruePositive}{Positive}$$

$$FPR = \frac{FalsePositive}{Negative}$$

$$Precision = \frac{TruePositive}{PredictedPositive}$$

2.5 ALS Use Case

To evaluate the effectiveness of our method, a case study for ALS disease (https://www.omim.org/entry/105400) was conducted. ALS is a neurodegenerative disorder characterized by the death of motor neurons in the brain and spinal cord, leading to skeletal muscle atrophy and eventually paralysis. With a prevalence rate of 5·2 per 100,000 individuals, and genetic heterogeneity with several genes responsible, it is a highly complex disease which is the subject of a great deal of research.

In order to find drugs that could be repurposed for ALS, we select all drugs that are not already linked to ALS in our KGs. Then for each drug, we link it to ALS, forming a new triple, create features as described previous and use the trained model to classify these triples as true or false.

Table 4. Accuracy results of our models for all three configurations of feature selection.

Method/Dataset	Oregano	Hetionet	BioKG
1-hop neighbors	0·973 (0·971–0·975)	0·910 (0·867–0·946)	0·984 (0·983–0·987)
2-hop neighbors	0·999 (0·999–1·000)	0·975 (0·953–0·996)	0·999 (0·998–0·999)
PageRank-neighbors	0·974 (0·972–0·976)	0·950 (0·913–0·978)	0·985 (0·983–0·986)

3 Results

3.1 Triple Classification

After creating the features, we obtained 60,990 positive triples and 609,900 negative triples for Oregano, 66,868 positive triples and 668,680 negative triples for BioKG, 483 positive triples and 4,830 negative triples for Hetionet, with the choice of ten negative triples for one positive triple yielding the best results (6). The number of features depends on the number of different types of nodes for each KG: 12 for Oregano, 6 for BioKG and 11 for Hetionet (??)

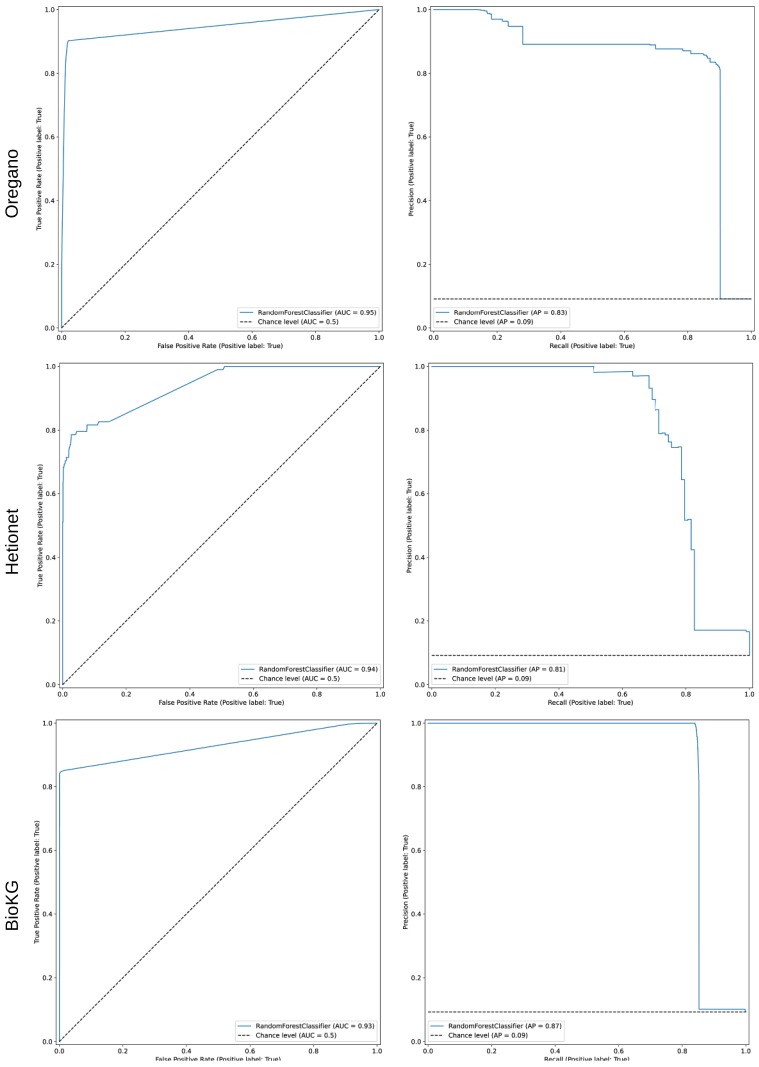


Fig. 2. Area under receiver-operator characteristic (left) and precision-recall (right) for the triple classification task on the three KGs. Top row are results for Oregano, middle row for Hetionet and last row for BioKG.

Table 5. Number of features, positive and negative samples constructed for each knowledge graph. The number of features is the same as the number of types of nodes in the KG.

Dataset	#Feature	#Positive	#Negative
Oregano	12	60,990	609,900
Hetionet	11	483	4,830
BioKG	6	66,868	668,680

Table 6. Median number of nodes retrieved by selecting neighbors using the three different methods.

Method/Dataset	Oregano	Hetionet	BioKG
1-hop neighbors	1	2	27
2-hop neighbors	574	6043	3375
PageRank-neighbors	50	22	48

From Table 4, we observe a high accuracy for all three methods of features selection. This first result validates the intuition that neighbor information is sufficient to train an RF classifier for the large-scale KG triple classification task. As we focus on explainable predictions, we then focus on interpreting the results of the model trained using PageRank-neighbors, as it allows us to control the amount of information contained in the features (best hyperparameters of these models can be found in ??). In this configuration, we computed the AUROC and the precision-recall results for the three KGs (Figs. 2, first column). We obtain AUROC values of 0·945 (95% CI 0·942–0·948) for Oregano, 0·939 (95% CI 0·911–0·968) for Hetionet and 0·925 (95% CI 0·922–0·928) for BioKG.

As the AUROC can be misleading for classifiers trained on imbalanced datasets, we then observe the precision-recall curves (Figs. 2, second column). For all three datasets, we can find a single probability threshold with both good precision and good recall. For Oregano, the model achieves a precision of 0·832 (95% CI 0·827–0·836) and a recall of 0·886 (95% CI 0·882–0·890), for Hetionet a precision of 0·923 (95% CI 0·894–0·963) and a recall of 0·743 (95% CI 0·687–0·799) and for BioKG a precision of 0·999 (95% CI 0·998–0·999) and a recall of 0·833 (95% CI 0·828–0·837). A high precision value indicates that our model very rarely classifies a false triple as a true one. We observe good precision scores for all three datasets, especially for Hetionet (0·923) and BioKG (0·999).

On the other hand, we observe recall values higher than 0·80 for Oregano and BioKG and slightly lower for Hetionet (0·743). Overall, our experiments show better performance in terms of precision than recall. In the case of DR, a high precision score is a more desirable property, as it indicates that if our model classifies a non-existent triple as true, this new triple is very likely to be of interest.

3.2 Explainability of Results

One of the advantages of our method is the possibility of systematically explaining the results. MHR models are also explainable, based on the paths used by the model to make the prediction. However, when running the two neuro-symbolic models MultiHop and SQUIRE for the link prediction task of the relationship *treats* [13,17], we found that when these models are able to provide results, the explanations are quite insufficient (7). The strength of our method lies in the fact that for each result, we can directly observe the features used to make the

prediction, *i.e.* the most important neighbors of the nodes forming the triple of interest. These neighbors can be found by simply querying a database containing the KG, such as Neo4j.

Table 7. Percentage of correct predictions with an explanation for the three MHR models.

Model/Dataset	Oregano	Hetionet	BioKG
AnyBURL	100	100	100
SQUIRE	8·4	0	0
MultiHop	99	75	0

3.3 Proposed Drugs for ALS

Table 8. All drug candidates found to be potential ALS treatments by our model, with scientific publications supporting the results.

Drug	Literature
Metformin	[25,26]
Rosiglitazone	[25,27]
Atorvastatin	[28,29]
Simvastatin	[29,30]
Fluorouracil	[31]
Paliperidone	[32,33]
Oxaliplatin	None
Clopidogrel	None
Epirubicin	None
Nicotine	None
Bevacizumab	None
Aspirin	None

After classifying all possible triples, including ALS, we obtain 12 drugs proposed for repurposing from both Oregano and BioKG (Table 8). As ALS is not present in Hetionet, no results were found using this KG. These compounds are the following: Metformin, Rosiglitazone, Atorvastatin, Clopidogrel, Epirubicin, Fluorouracil, Nicotine, Aspirin, Oxaliplatin, Simvastatin, Bevacizumab and Paliperidone. For each of these results, the small graph containing the neighbors that serve as features provides a visual explanation, as shown in Fig. 3 for the case of Metformin and Rosiglitazone.

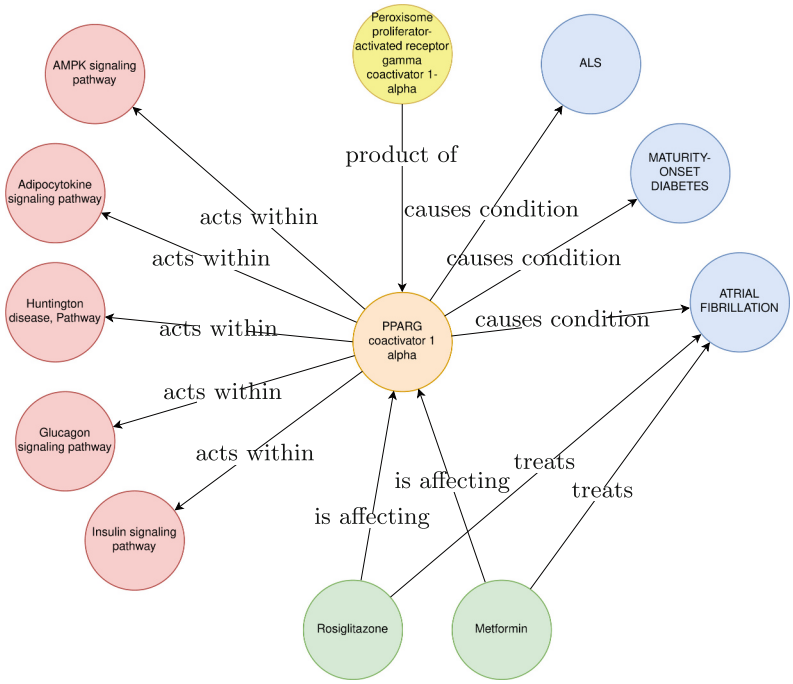


Fig. 3. An example of a graphic explanation supporting the findings that Metformin and Rosiglitazone are potential candidates for the treatment of ALS. Blue nodes are diseases, red are pathways, green are drugs, brown are genes, yellow are proteins. (Color figure online)

To evaluate the plausibility of our results, we searched for publications linking each drug to ALS using PubMed. Out of the 12 proposed drugs, we found that six were already being investigated for ALS, with a clinical trial underway for Metformin (https://clinicaltrials.gov/study/NCT04220021). As for Rosiglitazone, it has been shown that ALS patients had a reduction in skeletal muscle PPARG coactivator 1 alpha mRNA and protein [25], and that this drug interacts with the PPARG coactivator 1 alpha gene by increasing its expression [27]. Regarding Metformin and Rosiglitazone, Russel *et al.* showed that ALS patients had a reduction in skeletal muscle PPARG coactivator 1 alpha mRNA and protein [25]. In return, Muhlhausler and Aatsinki indicated that Rosiglitazone and Metformin increased the expression of the PPARG coactivator 1 alpha gene [26,27]. For Atorvastatin, Merwin *et al.* pointed out the link between ALS and a low PON1 gene activity [29], when Sardo *et al.* showed that Atorvastatin created an increased expression of PON1 mRNA and protein [28]. Similarly for Simvastatin, Deakin proved that it can modulate expression in vitro of the PON1 gene and is associated with increased serum paraoxonase concentration and activity [30]. Concerning Fluorouracil, Rando *et al.* published a study confirming that this anti-cancer drug was able to increase lifespan, delay disease

onset and improve motor performance in ALS mice [31]. Lastly, the publication of Aouti *et al.* suggested that the Paliperidone may avert SOD1 aggregation responsible of ALS, hence being used as a lead for ALS drug development [33].

4 Discussion

This study focuses on improving triple classification methods for the task of DR using only the neighborhood information of nodes in a KG. Experiments were conducted on three KGs, two of them being commonly used for DR [4,5,7,34,35]. Our method is the first to test the performance of an RF classifier on these KGs for DR, without using external information or KG embedding methods, hence focusing on explainable results. Although not all our results can be linked to scientific publications, the fact that half of them can be, with a drug in clinical trials, is strong evidence of the value of our findings. We look forward to further investigations of each of these drugs by ALS experts.

The first advantage of our method is its excellent performance, compared to the state-of-the-art, for the task of triple classification on large-scale biomedical KGs. Indeed, other alternative studies have shown the interest of RF for triple classification, but only on KGs built for a specific task [15,16,36]. Our method is suitable for large-scale KGs that have not been specifically constructed for DR. These KGs are built using information from different freely online data sources, such as UniProt, REACTOME, DrugBank, SIDER, PharmGKB. Working with KGs therefore provides potential DR results for numerous different drugs and diseases. The choice of a particular disease or drug depends on the interest of clinicians, but most known drugs and diseases can be found in these KGs, as they are intended to be comprehensive.

The second advantage of our method lies in its excellent results in terms of accuracy. For all three datasets, the model is able to clearly distinguish between positive and negative triples. Despite the construction of more negative triples than positive triples existing in the KGs, our method does not suffer from an overprediction of the negative class. This is an important behavior because the goal of DR is to provide researchers with a list of potential drugs for clinical trials, and only promising drugs. Our model presents a very low false positive rate on all three KGs (0·023 on Oregano, 0·055 on Hetionet, 0·004 on BioKG), indicating its ability to provide only insightful DR opportunities and very few "false alarms".

The third advantage is that our method provides fully transparent and explainable results. Works on using RF on KGs have been done using node and relation embeddings to create the features of the classifier [16,19,20]. In doing so, the interpretation of the most important features for a prediction is impossible, as the features themselves are not interpretable. In our case, we focus on creating features that can be found later in the KGs and understandable by clinicians. For each predicted triple, whether true or false, its most important neighbors for the prediction can easily be retrieved and provide meaningful information about related proteins, genes or biological pathways, allowing the prediction to be accepted or discarded based on clinicians' knowledge.

The fourth advantage of our method is its speed in obtaining results. In a previous study, an RF classifier for triple prediction with explainable features has been tested using semantic paths between drug and gene pairs [36]. This method provided explainable features in the form of existing paths in the KG and used them to train the classifier. The KG constructed for this work is a network of drugs and genes, representing approximately 620,000 triples. To build the features, the authors reported a runtime of 19·45 h using an Ubuntu server with 12 GB of RAM and 8 CPU cores. In our case, creating the BioKG features (more than two million triples) takes around 25 min with the same hardware configuration.

However, our study has certain limitations. First, unlike link prediction models, our method does not allow new triples to be scored easily and quickly. Link prediction models use the embeddings of each node and relation and a scoring function to determine the plausibility of a triple, which makes it possible to quickly score data that has never been seen before. In our case, after training the model, all the potential combinations of DRUG-DISEASE triples should be constructed. Only then, we can retrieve the features for each of these triples and use the trained model to classify them. Secondly, when used on a very sparse KG, our method may create uninformative explanations, because with very few connections, the neighbors of different triples may be the same, creating similar explanations for different triples. Conversely, for KGs whose nodes are extremely connected, some graphical explanations may lack readability, as the common neighbors selected to build the features may be numerous. In terms of the provided explanations, we lack a gold standard for assessing the quality of explanations. Even if, in our case, the explanations are based on real data contained in the KGs (*i.e.* the neighbors of the predicted triple), it might be interesting to see how these explanations are evaluated by clinicians. Finally, questions may be raised as to why we found results supported by scientific publications that were not present in the KGs used in this study. Most of the scientific articles supporting our results were published years ago, which rules out the hypothesis that the results were too recent to be integrated in the KGs. Moreover, some links between the proposed drugs and ALS can be found in the CTD database (https://ctdbase.org/), which is not included in the knowledge sources used to construct the KGs. This shows how difficult it is to build complete KGs, including all available knowledge on a given topic.

Declarations

Acknowledgments. This work was supported by DPH (17-EURE-0019), but it played no role in the design of the study, the collection and analysis of data, the decision to publish or the preparation of the manuscript. We thank the Translational Research in Neurogenetic Diseases (NRGEN) and the Glia-neuron Interactions teams for their help in the use case of ALS.

Author contributions. MD - Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review & editing.

FM - Conceptualization, Data curation, Formal analysis, Project administration, Resources, Supervision, Validation, Writing – review & editing.

AZ - Conceptualization, Supervision, Validation, Writing – review & editing.

GD - Conceptualization, Data curation, Formal analysis, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding Information. MD PhD thesis was supported by the French Ministry of Higher Education and Research. Further, he is attending a professional post-graduate training program supported by the EUR Digital Public Health Program, funded by the French National Agency of Research. The research work of GD is partly supported by the RHU TALENT project (France 2030 ANR Funding ANR-23-RHUS-0015), which is coordinated by CHU Bordeaux France.

Data Availability Statement. The complete code for training and validating the random forest is available at: https://gitub.u-bordeaux.fr/mdrance/random-forest-for-drug-repurposing. The MHR models used in this study are open source and available at:

https://github.com/bys0318/SQUIRE,

https://github.com/salesforce/MultiHopKG and

https://web.informatik.uni-mannheim.de/AnyBURL/. The three KGs we used are freely available and can be found at:

https://github.com/AstraZeneca/awesome-drug-discovery-knowledge-graphs/tree/master or on our repository https://gitub.u-bordeaux.fr/mdrance/random-forest-for-drug-repurposing.

Ethics Approval and Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

Competing Interests. We declare no competing interests.

References

1. DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: Innovation in the pharmaceutical industry: new estimates of RD costs. J. Health Econ. **47**, 20–33 (2016). https://doi.org/10.1016/j.jhealeco.2016.01.012
2. Wouters, O.J., McKee, M., Luyten, J.: Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA **323**, 844–853 (2020). https://doi.org/10.1001/jama.2020.1166
3. Morgan, P., et al.: Impact of a five-dimensional framework on RD productivity at Astrazeneca. Nat. Rev. Drug Discov. **17**, 167–181 (2018). https://doi.org/10.1038/nrd.2017.244
4. Walsh, B., Mohamed, S.K. Nováček, V.: Biokg: a knowledge graph for relational learning on biological data. In: Proceedings of the 29th ACM International Conference on Information Knowledge Management, CIKM '20, pp. 3173–3180. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3340531.3412776

5. Himmelstein, D.S., et al.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife **6**, e26726 (2017). https://doi.org/10.7554/eLife.26726
6. Zhang, R., et al.: Drug repurposing for covid-19 via knowledge graph completion. J. Biomed. Inf. **115**, 103696 (2021). https://www.sciencedirect.com/science/article/pii/S1532046421000253
7. Boudin, M., Diallo, G., Drancé, M., Mougin, F.: The oregano knowledge graph for computational drug repurposing. Sci. Data **10**, 871 (2023). https://doi.org/10.1038/s41597-023-02757-0
8. Lopes, P., et al.: Gathering and exploring scientific knowledge in pharmacovigilance. PLoS ONE **8(12)** (2013)
9. Djeddi, W.E., Hermi, K., Ben Yahia, S., Diallo, G.: Advancing drug-target interaction prediction: a comprehensive graph-based approach integrating knowledge graph embedding and protbert pretraining. BMC Bioinf. **24**, 488 (2023)
10. Xiao, Y., et al.: Repurposing non-pharmacological interventions for alzheimer's disease through link prediction on biomedical literature. Sci. Rep. **14** (2023)
11. Rudin, C., et al.: Interpretable machine learning: fundamental principles and 10 grand challenges. Stat. Surv. **16**, 1–85 (2022). https://doi.org/10.1214/21-SS133
12. Das, R., et al.: Go for a walk and arrive at the answer: reasoning over paths in knowledge bases using reinforcement learning. ArXiv **abs/1711.05851** (2017). https://arxiv.org/abs/1711.05851
13. Lin, X.V., Socher, R., Xiong, C.: Multi-hop knowledge graph reasoning with reward shaping. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3243–3253. Association for Computational Linguistics, Brussels (2018). https://doi.org/10.18653/v1/D18-1362
14. Meilicke, C., Chekol, M.W., Ruffinelli, D., Stuckenschmidt, H.: Anytime bottom-up rule learning for knowledge graph completion. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19 International Joint Conferences on Artificial Intelligence Organization, pp. 3137–3143 (2019). https://doi.org/10.24963/ijcai.2019/435
15. Olayan, R.S., Ashoor, H., Bajic, V.B.: DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. Bioinformatics **34**, 1164–1173 (2017). https://doi.org/10.1093/bioinformatics/btx731
16. Djeddi, W.E., Hermi, K., Ben Yahia, S., Diallo, G.: Advancing drug-target interaction prediction: a comprehensive graph-based approach integrating knowledge graph embedding and protbert pretraining. BMC Bioinf. **24**, 488 (2023). https://doi.org/10.1186/s12859-023-05593-6
17. Bai, Y., et al.: SQUIRE: a sequence-to-sequence framework for multi-hop knowledge graph reasoning. In: Goldberg, Y., Kozareva, Z. Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1649–1662. Association for Computational Linguistics, Abu Dhabi (2022). https://doi.org/10.18653/v1/2022.emnlp-main.10
18. Lv, X., et al.: Is multi-hop reasoning really explainable? towards benchmarking reasoning interpretability. In: Moens, M.-F., Huang, X., Specia, L. Yih, S. W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8899–8911. Association for Computational Linguistics, Online and Punta Cana (2021). https://doi.org/10.18653/v1/2021.emnlp-main.700

19. Zhao, B.-W., et al.: iGRLDTI: an improved graph representation learning method for predicting drug–target interactions over heterogeneous biological information network. Bioinformatics **39**, btad451 (2023). https://doi.org/10.1093/bioinformatics/btad451
20. Sousa, R.T., Silva, S., Pesquita, C.: Explaining protein-protein interactions with knowledge graph-based semantic similarity. Comput. Biol. Med. **170**, 108076 (2024). https://doi.org/10.1016/j.compbiomed.2024.108076
21. Malas, T.B., et al.: Drug prioritization using the semantic properties of a knowledge graph. Sci. Rep. **9**, 6281 (2019). https://doi.org/10.1038/s41598-019-42806-6
22. Qian, J., Li, G., Atkinson, K., Yue, Y.: Understanding negative sampling in knowledge graph embedding. Int. J. Artifi. Intell. Appl. **12**, 71–81 (2021). https://doi.org/10.5121/ijaia.2021.12105
23. Lei, D., et al. Learning collaborative agents with rule guidance for knowledge graph reasoning, 8541–8547 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.688
24. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking : Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, pp. 161–172 (1998). http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
25. Russell, A.P., et al.: Disruption of skeletal muscle mitochondrial network genes and mirnas in amyotrophic lateral sclerosis. Neurobiol. Dis. **49**, 107–117 (2013). https://doi.org/10.1016/j.nbd.2012.08.015
26. Aatsinki, S.-M., et al.: Metformin induces pgc-1α expression and selectively affects hepatic pgc-1α functions. Br. J. Pharmacol. **171**, 2351–2363 (2014). https://doi.org/10.1111/bph.12585
27. Muhlhausler, B.S., Morrison, J.L., McMillen, I.C.: Rosiglitazone increases the expression of peroxisome proliferator-activated receptor-γ target genes in adipose tissue, liver, and skeletal muscle in the sheep fetus in late gestation. Endocrinology **150**, 4287–4294 (2009). https://doi.org/10.1210/en.2009-0462
28. Sardo, M.A., et al.: Antioxidant effect of atorvastatin is independent of pon1 gene t(–107)c, q192r and l55m polymorphisms in hypercholesterolaemic patients. Curr. Med. Res. Opinion **21**, 777–784 (2005). https://doi.org/10.1185/030079905X45170. notePMID: 15969877
29. Merwin, S.J., Obis, T., Nunez, Y., Re, D.B.: Organophosphate neurotoxicity to the voluntary motor system on the trail of environment-caused amyotrophic lateral sclerosis: the known, the misknown, and the unknown. Arch. Toxicol. **91**, 2939–2952 (2017). https://doi.org/10.1007/s00204-016-1926-1
30. Deakin, S., Leviev, I., Guernier, S., James, R.W.: Simvastatin modulates expression of the pon1 gene and increases serum paraoxonase. Arterioscler. Thromb. Vasc. Biol. **23**, 2083–2089 (2003). https://doi.org/10.1161/01.ATV.0000096207.01487.36
31. Rando, A., et al.: Chemotherapeutic agent 5-fluorouracil increases survival of sod1 mouse model of ALS. PLoS One **14** (2019). https://doi.org/10.1371/journal.pone.0210752
32. Yu, M., et al.: Network medicine informed multi-omics integration identifies drug targets and repurposable medicines for amyotrophic lateral sclerosis. bioRxiv (2024). https://www.biorxiv.org/content/early/2024/03/30/2024.03.27.586949
33. Aouti, S., Padavattan, S., Padmanabhan, B.: Structure-based discovery of an antipsychotic drug, paliperidone, as a modulator of human superoxide dismutase 1: a potential therapeutic target in amyotrophic lateral sclerosis. Acta Crystallogr. Sect. D **79**, 531–544 (2023). https://doi.org/10.1107/S2059798323003649

34. Bonner, S., et al.: Understanding the performance of knowledge graph embeddings in drug discovery. Artif. Intell. Life Sci. **2**, 100036 (2022). https://doi.org/10.1016/j.ailsci.2022.100036
35. Bonner, S., et al.: A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. Briefings Bioinf. **23**, bbac404 (2022). https://doi.org/10.1093/bib/bbac404
36. Aisopos, F., Paliouras, G.: Comparing methods for drug-gene interaction prediction on the biomedical literature knowledge graph: performance versus explainability. BMC Bioinf. **24**, 272 (2023). https://doi.org/10.1186/s12859-023-05373-2

Computational Prediction of Intracellular Signaling Behavior via Machine Learning

Pamela Romero[1], Juliette Gourdon[1], Romain Yvinec[1,2], Frederic Jean-Alphonse[1,2], and Misbah Razzaq[1(✉)]

[1] INRAE, CNRS, Université de Tours, PRC, 37380 Nouzilly, France
{pamela.romero-jofre,juliette.gourdon,romain.yvinec,frederic.jean-alphonse,misbah.razzaq}@inrae.fr
[2] Inria, Inria Saclay-Ile-de-France, 91120 Palaiseau, France

Abstract. Assessing the dynamics of intracellular signaling processes under various conditions such as protein-protein interactions, protein conformational changes, dose-dependent drug effects, and ligand binding is crucial for biologists and pharmacologists. However, generating such data can be time-consuming and costly. In this study, we used data obtained by Bioluminescence Resonance Energy Transfer (BRET) in live cells to develop models to predict the temporal dynamics of intracellular second messenger cAMP (cyclic adenosine monophosphate). We formulated the task of predicting time series in a supervised manner and employed decision tree, random forest, and XGBoost models within a multiple-input, multiple-output framework, enabling simultaneous forecasting of multiple future time steps. Our findings demonstrate that our model achieves, in the main experiment, a mean absolute error of less than 0.018 on the testing set. This model is computationally efficient, requiring only ≈ 16 seconds for training, validation, and testing. Finally, we identified that BRET signal intensities are the most important feature for predictions among the set of features in the models. These results highlight the potential of machine learning models for advancing research in dynamical biological processes, particularly in cellular signaling and pharmacological systems.

Keywords: Machine learning · Random forest · Decision tree · Bioluminescence Resonance Energy Transfer (BRET) · Intracellular signalling · Time series

1 Introduction

Cell surface G-protein-coupled receptors (GPCRs) are a critical class of pharmacological targets, representing about 30% of FDA-approved drugs. The functional selectivity (biased signaling) of GPCRs is an important matter regarding pharmacological discovery, and is currently an active area of research with many statistical and mathematical developments [7,13,14,17,24]. In this study, we aimed to predict the cAMP (cyclic adenosine monophosphate) signaling

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 263–279, 2026.
https://doi.org/10.1007/978-3-032-20322-9_19

dynamics downstream of the luteinizing hormone receptor (LHR) and the follicle-stimulating hormone receptor (FSHR), two GPCRs essential for regulating reproductive physiology in mammals. cAMP is an intracellular second messenger, whose concentration is controlled by Gs activation, and the subsequently transmembrane adenylate cyclases (ADCY). cAMP dynamics are shaped by two main pathways: (i) the receptor activation by its hormone binding and the subsequently Gs/ADCY mediated cAMP concentration elevation, and (ii) the cAMP degradation by intracellular phosphodiesterases (PDEs). Bioluminescence Resonance Energy Transfer (BRET) is widely used to measure dynamic cellular events such as signaling, protein-protein interactions, ligand binding, and protein conformational changes in live cells. This technique thus allows gathering important information and knowledge regarding normal or pathological biological processes. To assess cAMP dynamics in live cells, we then used a cAMP BRET sensor transfected into cells expressing the Gs-coupled GPCRs of interest [17].

Time series prediction is a well-established area of research across various applications [23]. Time series prediction can be categorized by forecast horizon into single-step and multi-step forecasting. In single-step forecasting, a one-to-one prediction is made where the present input generates a single future output. In contrast, multi-step forecasting is feedback-based, incorporating prediction information from previous time steps. While this method requires less information, it may accumulate errors since it uses approximated inputs instead of actual values [1]. We adopt a multiple-input, multiple-output strategy that forecasts several future time steps simultaneously [21,22]. In this approach, the model is trained on an initial segment of time series data to predict subsequent segments. For simplicity, we refer to this as the "multi-step model". We focus on decision tree, random forest, and XGBoost models due to their flexibility in capturing non-linear relationships and complex interactions between variables like ligand, dose, receptor, and perturbation. These models are well-suited for predictions in dynamic biological systems, where data can be noisy and complex. The feature importance is calculated to understand how different biological factors influence the prediction of the model, such as receptor activation, Gs-coupled adenylate cyclase activity, or phosphodiesterase-mediated degradation of cAMP, which can ultimately help to understand the signaling components that affect the intracellular second messenger dynamics measured by the cAMP BRET sensor.

2 Data and Methods

2.1 Biological Protocol for Dataset Generation

HEK293A or mLTC-1 cells per well were seeded in 96-well plates previously coated with 0.01% polylysine. Cells were transiently transfected in suspension using Metafectene Pro transfection reagent (Biontex Laboratories, München, Germany) according to the manufacturer's instructions, using DNA quantities specified. Nb37 was transfected under the relevant experimental conditions. Forty-eight hours post-transfection, BRET measurements were conducted upon

addition of 5μM coelenterazine-H (Interchim) diluted in Ca^{2+}/Mg^{2+}-free PBS, containing various concentrations of ligand, and with the appropriate concentration of chemical drug for the concerned conditions. For experiments performed in the presence of Dyngo4a, PitStop2, ES9-17, or YM254890, cells were preincubated for 20–35 min in the presence of drug-containing buffers before the measurements were taken. For experiments performed in the presence of pertussis toxin (PTX), the compound was added to the cells 16 h prior to the experiment. Signals were recorded for at least 60 min with a Mithras LB 943 plate reader (Berthold Technologies GmbH & Co., Wildbad, Germany), and the BRET ratio (480 nm/540 nm) was calculated. At least four perturbations and eight different concentrations of ligand (one 96-well plate reader (to remove), including three replicates per condition) were measured simultaneously in a parallel manner. The overall cost of each experiment varies between €100 and €1,000 depending on the type of components used, and it is mainly associated with several factors: i) use of the plate reader; ii) cell culture and consumables (including ligand, drugs, and transfection reagents); iii) and human resources needed to run the experiments and data analysis. The final dataset is not yet public.

2.2 Dataset Processing

Our dataset includes temporal changes in ligand-induced cAMP BRET signals under different conditions. A time series signal consists of continuous values, and a time series with k data points can be represented as:

$$T_c = (t_c^1, t_c^2, \ldots, t_c^k) \tag{1}$$

where T_c represents the time series for a given condition c, and c denotes a specific condition determined by the dose, ligand, receptor, and perturbation. The dose refers to the concentration of the ligand in nanomolar units (nM). The dataset includes several dose values: 0, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100 nM. Ligands are molecules that bind to receptors. Our dataset includes three types of ligands: hCG, LH, and FSH [7]. We also have five types of receptors: hLHR, hLHR-T (a truncated version of hLHR missing 17 amino acids in the distal C-terminus, which affects its endocytic and post-endocytic trafficking [12]), mLHR, hFSHR, and mFSHR. These receptors are from two species (human and mouse) and are associated with three main hormones: hCG and LH for LHR, and FSH for FSHR. The final category, perturbations, includes pharmacological compounds (Dyngo4a, PitStop2, ES9-17, YM254890) that alter signaling pathways, or genomic modifications that allow for measuring cAMP at specific sub-cellular locations (we do not mix compounds in the design of experiments). mCherry was used as a control for compartmentalized conditions. Drugs (e.g., Dyngo4a, PitStop2) inhibit the enzymatic activity of different proteins involved in the endocytosis machinery or G protein signaling. We use "the optimal" dose concentration to inhibit the activity of their target, as reported in the literature.

Each BRET signal time series corresponds to a case with varying features: time, BRET signal, and conditions. The dataset is obtained from different biological experiments and includes three replicates of each case. To mitigate the

potential bias and data leakage caused by the replicates, we applied the average of the replicates for each case. We then divided the averaged replicates into separate sets. We performed a main experiment and a five-fold cross-validation. In the main experiment, after the average of the replicates, we divided the dataset into three subsets: (i) a training set for model training, (ii) a validation set for hyperparameter tuning, and (iii) a testing set to evaluate the model with held out data that is not included in either the training set or the validation set. The dataset contains 330 training cases (70% of tha data), 47 validation cases (10% of tha data), and 95 test cases (20% of the data). Each case has 64 data points over time and several different conditions. Moreover, each input case to the model is formed by 84 features: one value of dose (position 0 of the feature vector), four features of the ligand (positions 1 to 4 of the feature vector), six features of the receptor (positions 5 to 10 of the feature vector), nine features of the perturbation (positions 11 to 19 of the feature vector), 32 time points (positions 20 to 51 of the feature vector), and 32 data points (positions 52 to 83 of the feature vector). The time points in the dataset increase approximately by one to two-minute intervals, ranging from 0 to a maximum of 69 min on average.

In the five-fold cross-validation experiment, the complete dataset was partitioned into five subsets, each subset used once as the test set (≈ 95) and the remaining four for training (≈ 378). In this experiment, we did not use validation data. We used the same hyperparameter values found in the main experiment. Figure 1 illustrates the composition of the datasets used in the main and cross-validation experiments.

Normalization: The measurements vary across different ranges. Min-max normalization scheme scales the measurements to a common range, specifically between 0 and 1, to eliminate bias toward larger or smaller values [5]. Equation 2 is used to perform normalization in this study:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

where x denotes an original value and x_{norm} is the normalized value, $\max(x)$ and $\min(x)$ are the maximum and minimum values of the input data. In this study, both the BRET signal and the dose are normalized to avoid introducing bias. The normalized signals retain the aspect of the original signals.

Transformation of Different Conditions: We applied different transformation schemes, using one-hot encoding for ligands, receptors, and perturbations to avoid introducing ordinal relationships between categories [25]. The ligand variable has four possible values. With one-hot encoding, each input is mapped to a 4-dimensional vector, with a 1 in the position corresponding to the input value and 0 s in the others. The same scheme is applied to both receptors and perturbations.

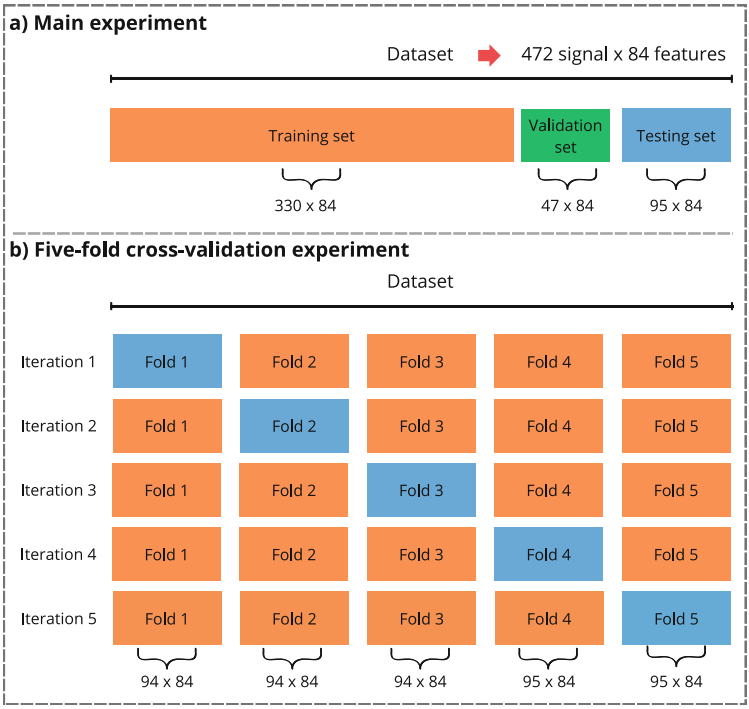


Fig. 1. Training (orange color), validation (green color), and testing sets (blue color) for the main and cross-validation experiments. (Color figure online)

2.3 Predicting Signalling Processes Using Machine Learning Models

In this work, we address the task of time series prediction using decision tree, random forest, and XGBoost algorithms. These models are all tree-based but differ in their complexity and underlying techniques. The decision tree is the simplest, while XGBoost is a more advanced method. Random forest employs bagging [2], an ensemble technique that aggregates multiple decision trees to improve robustness and reduce overfitting. In contrast, XGBoost [6] (Extreme Gradient Boosting) uses boosting [8], a sequential ensemble method that builds trees iteratively to minimize a loss function using gradient descent optimization [9]. Random forest is widely used for regression tasks due to its robustness. A decision tree makes predictions by learning a set of rules from the training data [11]. Gradient boosting trees, like XGBoost, combine multiple weak learners to form a strong predictive model. Tree-based methods are simple and practical for interpretation. In the research of Grinsztajn et al. [10], the performance of various tree-based models and deep learning models was evaluated using 45 tabular datasets. As a result, tree-based models yield more accurate predictions for these tabular datasets.

In this research, we developed a multi-step prediction model to forecast multiple future time steps of the signal simultaneously. Given a fixed number of features, the model predicts not only the next time step but also several subsequent steps. To formulate the time series prediction task, we transformed the sequential signal data into a supervised learning problem by using past observations as input features to predict future time points. Specifically, the input to the model consists of the first 50% of the signal from each time series (32 data points over time), and the model predicts the remaining 50% (32 data points following) (see Fig. 2). This approach requires all cases to have the same length of sequences.

Time	**BRET**	**Prediction P = 50%**
T_0	x_0	
T_1	x_1	
T_2	x_2	
T_3	x_3	
T_4		$\hat{x}_4$
T_5		$\hat{x}_5$
T_6		$\hat{x}_6$
T_7		$\hat{x}_7$

Fig. 2. An example of the multi-step model predicting future time steps of the signals is shown. The blue bounding box contains the input to the model (T_0 to T_3 represents time steps and x_0 to x_3 are the signal values of these time steps), while the red bounding box represents the predicted signal ($\hat{x}_4$ to $\hat{x}_7$). T_4 to T_7 symbolizes the future time steps. (Color figure online)

2.4 Configuration of the Hyperparameters

This study uses the grid search algorithm from the scikit-learn library to determine the optimal configuration of hyperparameters, including the maximum depth of the tree, the minimum number of samples required to be at a leaf node, the number of trees in the forest, and the learning rate. In the decision tree model, the hyperparameters found were: `max_depth = 10`, `min_samples_leaf = 1`. In the random forest model, the optimal hyperparameters were: `max_depth = 10`, `min_samples_leaf = 1`, `n_estimators = 10`. We selected the following hyperparameters for the XGBoost model: `learning_rate = 0.5`, `max_depth = 3`, and `n_estimators = 10`.

2.5 Performance Evaluation

We use three metrics to evaluate the performance of our machine learning models: Mean Absolute Error (MAE), Area Under the Curve (AUC), and time-weighted Root Mean Square Error (RMSE). MAE quantifies the average absolute difference between the actual and predicted values of the time series and is calculated as shown in Eq. 3. The time-weighted RMSE applies a weighting factor that increases with time to emphasize errors occurring later in the series, as defined in Eq. 4. This weighting reflects the biological importance of accurately predicting later stages of the signal dynamics. In this study, the weights increase approximately linearly from 0 at the start of the time series to a maximum value near 1 at the final time point. For both MAE and time-weighted RMSE, an ideal value is zero, indicating perfect predictions.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where n is the total number of data points, y_i is the observed value, and $\hat{y}_i$ represents the predicted value.

$$\text{Time-weighted RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{w_i(y_i - \hat{y}_i)^2}{\sum w_i}} \tag{4}$$

where n, y_i, and $\hat{y}_i$ have the same meanings as in Eq. 3, and w_i denotes the weighting factor.

The Area Under the Curve (AUC) is calculated using the Trapezoidal rule [19]. The predicted and ground truth AUC values represent the total area under their respective curves plotted over time. Ideally, these values should be as close as possible, indicating that the model accurately captures the overall temporal dynamics of the signal. The AUC is computed according to Eq. 5:

$$\text{AUC} = \frac{1}{2}\sum_{i=1}^{n-1}(t_{i+1} - t_i)\,(\mathring{y}_i + \mathring{y}_{i+1})\,, \tag{5}$$

where n is the total number of data points, t_i denotes the time at the ith point, and $\mathring{y}_i$ represents the time series value at t_i, which can be either the observed value y_i or the predicted value $\hat{y}_i$.

2.6 Feature Importance

We employ three methods to study feature importance: mean decrease in impurity (MDI) [3], the permutation feature importance technique [3], and SHapley Additive exPlanations (SHAP) values [16]. MDI uses statistics from the training dataset to compute feature importance; however, these features may not necessarily be informative on the testing dataset. The permutation method addresses this issue, as it can be computed on unseen data [18]. SHAP values are a model-agnostic method based on game theory that quantify the contribution of each feature to an individual prediction.

3 Results

3.1 Comparative Models

We trained a decision tree, random forest, and XGBoost models using multiple feature vectors and their targets. The implementation of the model and the processing of the data is performed using the scikit-learn library [18]. The codes are available at this link: GitHub.

In Figs. 3, 4 and 5, we illustrate the predictions of our models on the testing set of the main experiment; the solid lines show the ground truth, the square figure shows the decision tree prediction, the triangle figure shows the random forest prediction, and the star figure shows the XGBoost prediction. The results of this method are satisfactory, with the predicted BRET signal values closely matching the ground truth. In Figs. 3, and 4, we present four cases with hCG ligand and a ligand concentration of 3 nM. Figure 3 shows the LHR receptor, while Fig. 4 shows the LHR-T receptor (truncated version of the LHR receptor). The red signal in both figures is the signal without perturbation, while the blue signal is the signal with Dyngo4a perturbation. Blocking receptor endocytosis using Dyngo4a results in a decrease in the cAMP signal, and the signal is now located at the bottom of the graph.

Figure 3 shows that in the absence of perturbation, random forest produces predictions closest to the ground truth, followed by XGBoost, while the decision tree shows greater deviation. Under perturbation, random forest maintains the highest accuracy, with the decision tree performing better than XGBoost, whose predictions diverge more significantly from the ground truth. Figure 4 shows that, consistent with Fig. 3, the best prediction performance in the absence of perturbation is achieved by the random forest model, followed by XGBoost, with the decision tree presenting the least accurate predictions. However, under perturbation, the decision tree model produces predictions closest to the ground truth, with random forest in second place and XGBoost diverging further. In both Fig. 3 and 4, XGBoost delivers predictions that tend to capture the overall trend (whether ascending or descending) of the signal, even when its absolute accuracy is lower than the other models. Furthermore, the decision tree model presents predictions with a more linear behavior.

In Fig. 5, we present four cases involving the hCG ligand and the LHR-T receptor. There are different ligand concentrations, which are represented by different colors: purple for 1 nM, red for 3 nM, green for 10 nM, and yellow for 30 nM. As the ligand dose increases, the signal values also increase. The prediction results show that both the random forest model and the XGBoost model tend to yield predictions that are very close to the ground truth and to each other. For example, at a dose of 10 nM, the predictions of both models align so closely that it is difficult to distinguish between them. Meanwhile, the predictions of the decision tree model are slightly further from the ground truth and more linear in comparison with the other models. However, it can also be observed that when the dose is equal to 10 nM, the decision tree model tries to capture the trends in the signals.

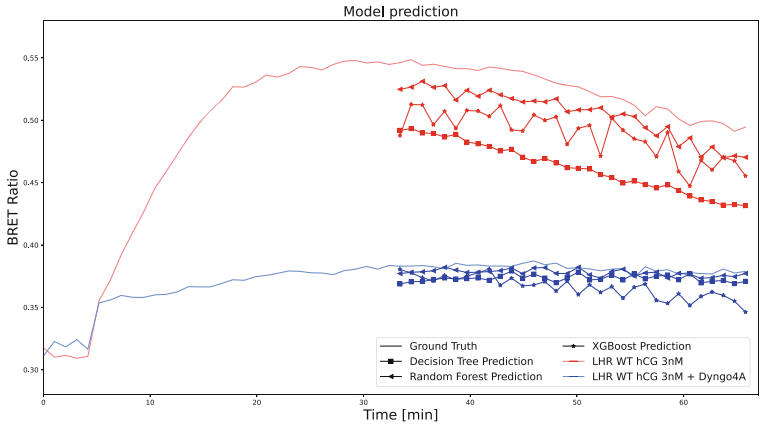


Fig. 3. Model prediction. Solid lines show the ground truth, the square figure shows the decision tree prediction, the triangle figure shows the random forest prediction, and the star figure shows the XGBoost prediction. (Color figure online)

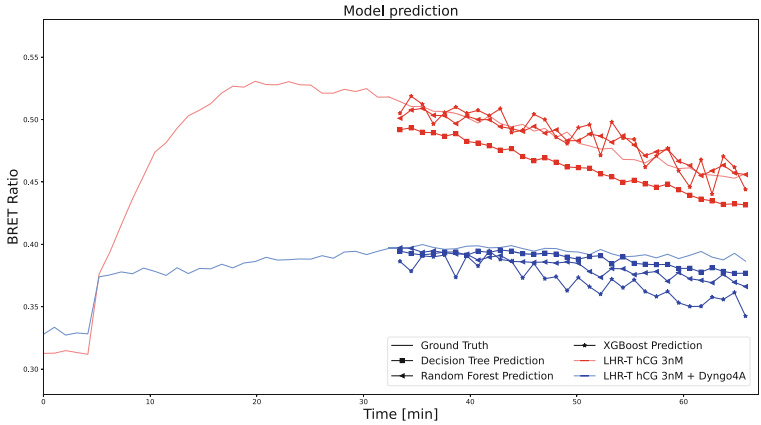


Fig. 4. Model prediction. Solid lines show the ground truth, the square figure shows the decision tree prediction, the triangle figure shows the random forest prediction, and the star figure shows the XGBoost prediction. (Color figure online)

In conclusion, the qualitative results show that the prediction using the random forest model tends to be close to the ground truth. The XGBoost model attempts to capture the trend of the signal. In contrast, the decision tree model has predictions with a more linear behavior.

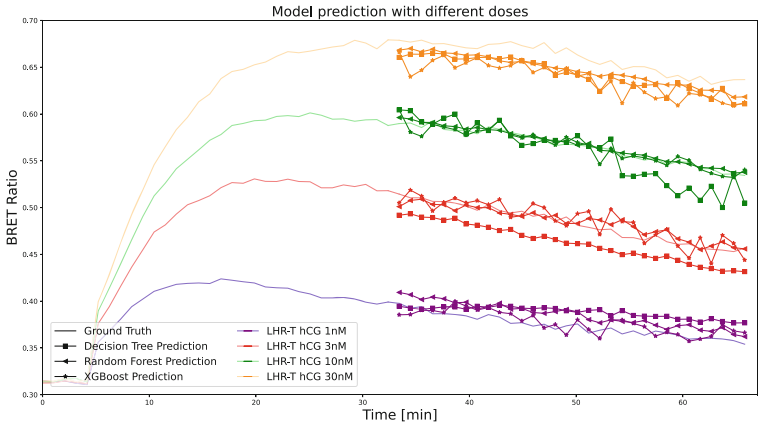


Fig. 5. Model prediction with different doses. Solid lines show the ground truth, the square figure shows the decision tree prediction, the triangle figure shows the random forest prediction, and the star figure shows the XGBoost prediction. (Color figure online)

3.2 Evaluation of Prediction Accuracy

Figures 6 and 7 present the main experimental results for the evaluation metrics, with colors indicating the different dataset partitions: blue for training, green for validation, and red for testing. Figure 6 illustrates the MAE results, where the random forest model achieved values of 0.0020, 0.0128, and 0.0140 for the training, validation, and testing sets, respectively. The decision tree model obtained MAE values of 0.0003, 0.0138, and 0.0158, while the XGBoost model yielded 0.0019, 0.0132, and 0.0173 in the respective sets. These errors are close to zero, with the lowest errors generally observed on the training set. Errors increase slightly in the validation and testing sets, but the differences across the sets are minimal. Figure 7 shows the results of the time-weighted RMSE, which was approximately 0 across all sets and models. Moreover, the three models produced very similar results in these evaluation metrics.

Figure 8 presents the AUC results for the prediction models. The decision tree and random forest models show comparable AUCs across the training, validation, and testing sets, with predicted and ground truth AUC values closely aligned. The XGBoost model exhibits a slightly larger gap between predicted and ground truth AUCs in the validation and test sets, which may reflect its higher sensitivity to data and hyperparameters.

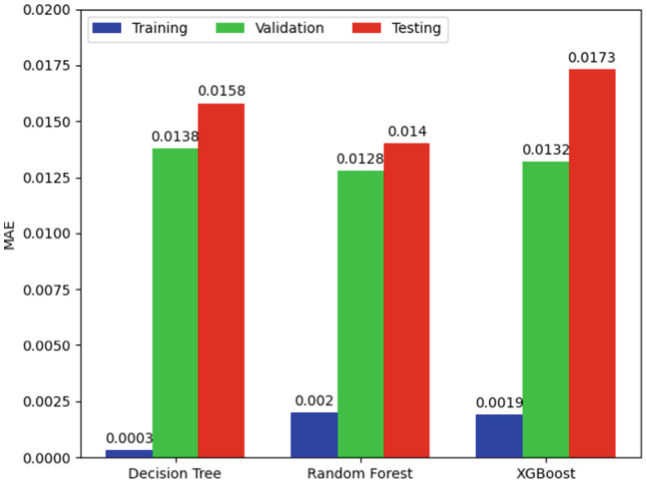


Fig. 6. MAE results. (Color figure online)

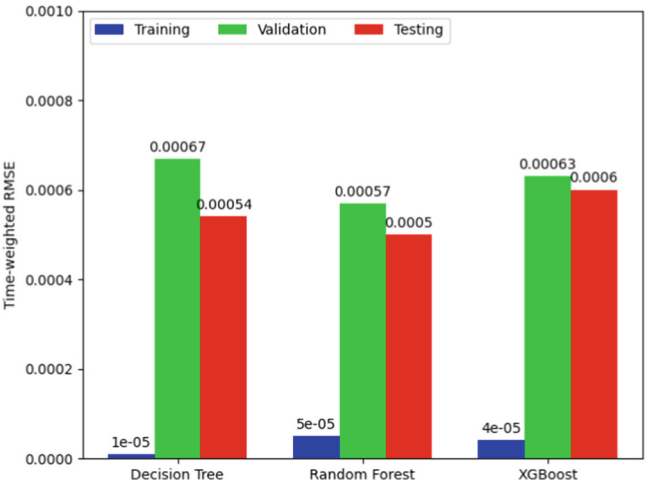


Fig. 7. Time-weighted RMSE results. (Color figure online)

In conclusion, in the quantitative results, the random forest model presents better results in the different metrics, followed by the decision tree model, and finally XGBoost. In summary, these results demonstrate that the multi-step prediction approach is capable of accurately forecasting several future steps based on initial signal values, as evidenced by the MAE, AUC and time-weighted RMSE resutls discussed above.

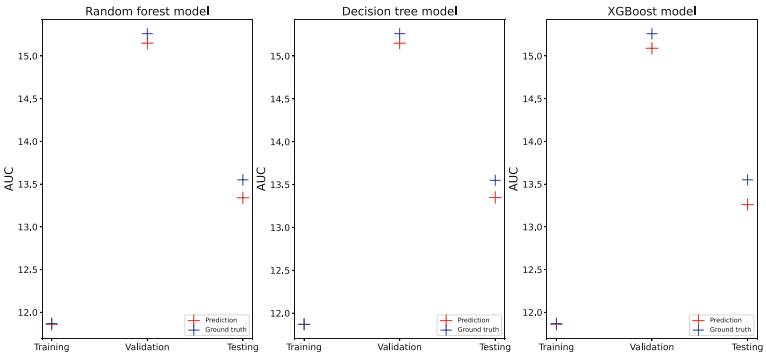


Fig. 8. Main experiment AUC results. Red represents the predicted signal, and blue represents the ground truth signal. (Color figure online)

3.3 Model Performance Under Cross-Validation

Tables 1, 2 and Fig. 9 show the cross-validation results on the test sets. The average MAE values (Table 1) were 0.026 for the random forest, 0.027 for both the decision tree and the XGBoost models. The time-weighted RMSE (Table 2) was approximately zero for all three models across all folds. The average AUC values (Fig. 9) were similar between the predictions and ground truth across the models. However, fold three exhibited higher error values across all three metrics compared to the other folds, likely reflecting differences in the test set conditions and time series data compared to the training sets. Overall, the error values in all folds were close to zero, which is ideal, and the differences between predicted and ground truth AUCs were minimal. The multi-step prediction approach remains efficient under cross-validation, requiring less than 16 s for training, validation, and testing for each model. Furthermore, the performance of the model under cross-validation is consistent with that observed in the main experiment. Among the models, random forest achieved the best average performance, followed by decision tree and then XGBoost.

Table 1. MAE results in cross-validation.

Folds	Decision Tree	Random Forest	XGBoost
Fold 1	0.0175	0.0154	0.0175
Fold 2	0.0002	0.0002	0.0005
Fold 3	0.0936	0.0937	0.0929
Fold 4	0.0092	0.0092	0.0097
Fold 5	0.0166	0.0134	0.0158
Average	0.027	0.026	0.027

Table 2. Time-weighted RMSE results in cross-validation.

Folds	Decision Tree	Random Forest	XGBoost
Fold 1	0.00072	0.00066	0.00080
Fold 2	0.00000	0.00000	0.00002
Fold 3	0.00253	0.00252	0.00250
Fold 4	0.00030	0.00029	0.00033
Fold 5	0.00056	0.00049	0.00055
Average	0.00082	0.00079	0.00084

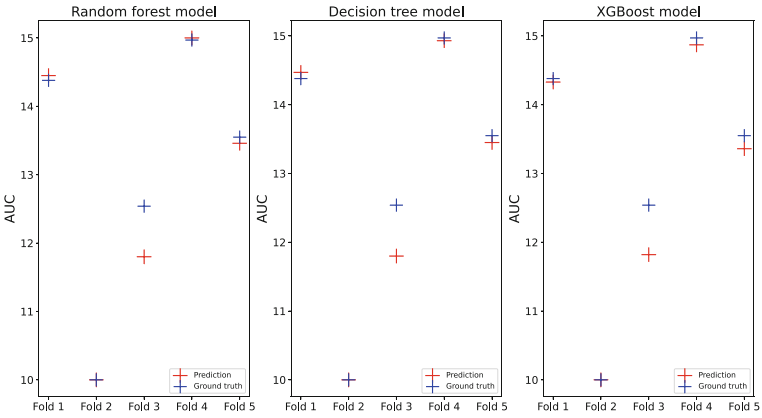


Fig. 9. Cross-validation average AUC results. Red represents the predicted signal, and blue represents the ground truth signal. (Color figure online)

3.4 Importance of Biological Factors

According to SHAP values, the last four BRET signal values had the most significant impact on predictions in the random forest model, whereas in the decision tree model, the last two values (at positions 82 and 83 of the feature vector) of the sequence were most influential. In the XGBoost model, the last five values of the BRET signal were the most critical for prediction. Additionally, the XGBoost model showed greater variability in SHAP values, with the dose variable exhibiting slightly higher importance than the decision tree and random forest models. This diversity of important features in XGBoost reflects its more complex boosting architecture.

According to the MDI feature importance, the decision tree model prioritizes the last two values of the BRET signal, the random forest emphasizes the last four values of the sequence, and the XGBoost model highlights the last five BRET signal values as most important. In the permutation technique, the decision tree presents results similar to MDI, whereas the random forest model highlights the last two values of the BRET signal. In the XGBoost model, the

permutation method suggests that the last BRET signal value is the most important feature, showing a modest shift from the MDI results.

These differences in importance likely stem from the distinct methodologies and biases inherent in different methods. All three explanation methods consistently indicate that later BRET signal values play the most critical role, while biological conditions (ligand, receptor, perturbation) did not play a significant role in any model. Meanwhile, dose was the only condition identified as slightly important using the SHAP method with the XGBoost model. The irrelevance of conditions in the case of MDI may also result from the inflation of the importance of numerical features (i.e., features with high cardinality). Permutation importance can also be biased when features are correlated, as permuting one may not eliminate its information, which can still be inferred from others, leading to underestimation [18]. Although SHAP values minimally capture feature contributions of dose, they ignore causality, which is contradictory in our case since the biological conditions define the signal. Further investigation is necessary to understand the potential contribution of biological conditions under different experimental settings.

3.5 Comparison with Traditional Models

Traditionally, dynamic models are used to predict signal behavior, especially for GPCR data. To build these models, a mathematical framework must be chosen, such as continuous-time Markov chains (for stochastic dynamics) or ordinary differential equations (ODEs) (for deterministic dynamics) [24]. For example, Bridge et al. [4] used ODEs to model cAMP signaling in response to GPCR ligands. They showed that ligand and receptor concentrations strongly influence cAMP responses and derived ligand-specific kinetic parameters. Similarly, Leelawattanachai et al. [15] modeled GPCR dynamics using ODEs to predict drugs affect on G protein activation, with estimated parameters matching well with experimental observations.

Traditional models face several challenges compared to machine learning approaches. A major issue is parameter identification: parameter values taken from prior work may not reflect the current biological context, leading to errors. Computational costs also increases, particularly during parameter fitting and simulation, as the size and complexity of the network increase [20]. Moreover, ODE-based models are often tailored to specific experimental setups, making cross-study comparison difficult. Finally, while decision tree-based models can handle high-dimensional data through feature selection, they do not capture causal mechanisms. However, they provide a scalable, data-driven way to predict complex signaling dynamics such as cAMP temporal patterns. In this study, we did not perform empirical comparisons with traditional dynamic models, as constructing and calibrating such models requires extensive parameter identification for each experimental condition, which was beyond the scope of this work. Furthermore, deep learning approaches were not pursued due to the limited size of the available dataset.

Finally, to compare the performance of tree-based models, we decided to use the linear regression model as a baseline. Linear regression exhibited a discrepancy in the quantitative results, particularly in cross-validation, compared to tree-based methods. For instance, in the time-weighted RMSE average results, linear regression obtained the highest value, equal to 0.00091, compared to the other models. In the average AUC results obtained, linear regression also yields predictions that are most distant from the ground truth. In terms of feature importance, time emerges as an essential feature for prediction, in contrast to the tree-based models.

4 Conclusion

In this work, we presented a machine learning approach to predict the dynamics of intracellular signaling processes. We employed decision tree, random forest, and XGBoost models in combination with a multi-step forecasting strategy. The proposed model effectively captures long-term dependency patterns and achieves a testing mean absolute error of less than 0.018.

Although initially applied to cAMP dynamics, it is quite versatile and can be adopted to predict dynamics of other intracellular signaling pathways, such as PKA, ERK, AKT, other second messengers production and degradation dynamics (i.e. calcium), recruitment of effector proteins (G proteins and β-arrestins in the case of GPCRs), maybe also ligand binding to a target. It provides a practical framework for characterizing GPCR signaling dynamics and offers potential applications in drug discovery by aiding the characterization of new therapeutic compounds.

To expand this methodology, we plan to incorporate additional biological conditions, including different types of ligands, genomic perturbations affecting receptor trafficking and signaling. It will help to understand receptor activity mechanisms and the dynamic signaling networks governing specific biological responses. A significant next step would be to fine-tune the model and test it with data from other laboratories. Future work will focus on validating model predictions through laboratory experiments and developing a user-friendly interface for biologists to generate and validate predicted time courses under various experimental conditions. Additionally, we aim to overcome the current limitation of our multi-step model, i.e., the fixed length of time series, by exploring deep learning approaches capable of handling variable-length sequences. This future direction will require collecting additional experimental datasets to train and evaluate such models sufficiently. Further, we will focus on performing a robustness analysis for the cross-validation experiment. Finally, we plan to investigate moving average normalization and improve the modeling approach by incorporating local trend information within the time series.

Acknowledgments. We acknowledge the financial support from the Pontificia Universidad Católica de Chile, iHEALTH, Inria, INRAE, MAbImprove Labex, CNRS, Le Studium Loire Valley, the Bill & Melinda gates foundation CONTRABODY grant.

Disclosure of Interests. The authors declare that there is no conflict of interest.

References

1. Bao, Y., Xiong, T., Hu, Z.: Multi-step-ahead time series prediction using multiple-output support vector regression. Neurocomputing **129**, 482–493 (2014)
2. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
3. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
4. Bridge, L., Chen, S., Jones, B.: Computational modelling of dynamic camp responses to gpcr agonists for exploration of glp-1r ligand effects in pancreatic β-cells and neurons. Cell. Signal. **119**, 111153 (2024)
5. Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., J. Tallón-Ballesteros, A.: The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis. In: International Conference on Soft Computing Models in Industrial and Environmental Applications, pp. 344–353. Springer (2023)
6. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
7. De Pascali, F., et al.: Pharmacological characterization of low molecular weight biased agonists at the follicle stimulating hormone receptor. Int. J. Mol. Sci. **22**(18), 9850 (2021)
8. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
9. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232 (2001)
10. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural. Inf. Process. Syst. **35**, 507–520 (2022)
11. James, G., Witten, D., Hastie, T., Tibshirani, R., et al.: An introduction to statistical learning, vol. 112. Springer (2013)
12. Jean-Alphonse, F., Bowersox, S., Chen, S., Beard, G., Puthenveedu, M.A., Hanyaloglu, A.C.: Spatially restricted g protein-coupled receptor activity via divergent endocytic compartments. J. Biol. Chem. **289**(7), 3960–3977 (2014)
13. Kenakin, T.: Biased receptor signaling in drug discovery. Pharmacol. Rev. **71**(2), 267–315 (2019)
14. Landomiel, F., et al.: Biased signaling and allosteric modulation at the fshr. Front. Endocrinol. **10**, 447467 (2019)
15. Leelawattanachai, J., Modchang, C., Triampo, W., Triampo, D., Lenbury, Y.: Modeling and genetic algorithm optimization of early events in signal transduction via dynamics of g-protein-coupled receptors: internalization consideration. Appl. Math. Comput. **207**(2), 528–544 (2009)
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30** (2017)
17. Namkung, Y., et al.: Quantifying biased signaling in GPCRs using BRET-based biosensors. Methods **92**, 5–10 (2016)
18. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

19. Purves, R.D.: Optimum numerical integration methods for estimation of area-under-the-curve (AUC) and area-under-the-moment-curve (AUMC). J. Pharmacokinet. Biopharm. **20**(3), 211–226 (1992)
20. Reiter, E., Yvinec, R., Crépieux, P., Poupon, A.: Coupling of recognition and effect in GPRC signaling computational modelling approaches. Encyclopedia of the molecular life sciences. In press, London (2016)
21. Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A.: A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. Expert Syst. Appl. **39**(8), 7067–7083 (2012)
22. Taieb, S.B., Sorjamaa, A., Bontempi, G.: Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing **73**(10–12), 1950–1957 (2010)
23. Tyralis, H., Papacharalampous, G.: Variable selection in time series forecasting using random forests. Algorithms **10**(4), 114 (2017)
24. Yvinec, R., Ayoub, M.A., De Pascali, F., Crépieux, P., Reiter, E., Poupon, A.: Workflow description to dynamically model β-arrestin signaling networks. Beta-Arrestins: Methods Protocols, 195–215 (2019)
25. Zheng, A., Casari, A.: Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc." (2018)

MeSH Concept Relevance and Knowledge Evolution: A Data-Driven Perspective

Jenny Copara[1(✉)], Nona Naderi[2], Gilles Falquet[1], and Douglas Teodoro[1]

[1] University of Geneva, Geneva, Switzerland
jenny.copara@etu.unige.ch, {gilles.falquet,douglas.teodoro}@unige.ch
[2] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, París, France
nona.naderi@universite-paris-saclay.fr

Abstract. The Medical Subject Headings (MeSH), one of the main knowledge organization systems in the biomedical domain, continuously evolves to reflect the latest scientific discoveries in health and life sciences. Previous research has focused on quantifying information in MeSH primarily through its hierarchical structure. In this work, we propose a data-driven approach based on information theory and network analysis to quantify the relevance of MeSH concepts. Our method leverages article annotations and their citation networks to compute four aspects of relevance—informativeness, usefulness, disruptiveness, and influence—over time. Using both the citation network and the MeSH hierarchy, we compute these relevance aspects and apply an aggregation algorithm to propagate scores to parent nodes. We evaluated our approach on MeSH terminology changes and showed that it effectively captures the evolution of concepts. The mean relevance of evolving concepts is higher compared to concepts that remained unchanged ($2.09E-03$ *vs.* $8.46E-04$). Moreover, we validated the framework by analyzing retracted articles and found that concepts used to annotate retracted articles (mean relevance: 0.17) differ substantially from those annotating non-retracted ones (mean relevance: 0.15). Overall, the proposed framework provides an effective method for ranking concept relevance and can support the maintenance of evolving knowledge organization systems.

Keywords: terminology evolution · concept relevance · information theory · network analysis

1 Introduction

The vast amount and complexity of the scientific biomedical literature require sophisticated methods of structuring and reasoning of knowledge [23,42]. Knowledge organization systems (KOSs)—including terminologies, thesauri, and ontologies—are designed to formally encode concepts, enhancing discoverability, representation, and reasoning across biomedical corpora [23]. Among these

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 280–299, 2026.
https://doi.org/10.1007/978-3-032-20322-9_20

KOSs, the Medical Subject Headings (MeSH) plays a central role by indexing, structuring, and facilitating search within the leading biomedical corpus, MEDLINE, which is accessible through the PubMed interface [26].

MeSH is a controlled vocabulary maintained by the U.S. National Library of Medicine [27]. Functionally, it is a thesaurus encompassing several types of concepts [1,23,35]. The primary units for indexing and retrieval are descriptors, which are used to annotate the MEDLINE/PubMed dataset. Additional elements, such as qualifiers and y concept records, provide further specification and context for descriptors. MeSH headings represent biomedical concepts and serve as standardized interpretations for groups of related medical terms [23,35]. MeSH headings are organized hierarchically into sixteen main categories, and descriptors map directly to their appropriate positions in the hierarchy [23]. Beyond facilitating search and question answering in the biomedical literature [17,18], MeSH also supports the study of structural and evolutionary patterns in biomedicine. Furthermore, this thesaurus assists in the evaluation of research grants and helps identify emerging research topics of high interest [14,22,43].

As knowledge in the biomedical domain evolves, MeSH needs to be periodically modified and updated to reflect these changes [28,32]. Based on the current knowledge state, MeSH concepts (i.e., headings) encode a certain amount of information, which may vary as knowledge develops [2]. The relevance of some concepts may decrease or increase, rendering them obsolete, necessitating drastic changes or additions [1,2,6]. As a consequence, the hierarchy structure can change. Thus, it is crucial to identify the most representative MeSH concepts [23,25] to summarize the content of the thesaurus, prioritize maintenance tasks in the KOS, and improve the performance of information retrieval systems that rely on this KOS, such as PubMed [5,6,12].

Researchers have applied different approaches to quantify the information encoded in MeSH by analyzing features such as the hierarchy, bibliometric indices of articles, and the co-occurrence of descriptors [1,2,7,22,24]. Balogh *et al.* [2] showed that nodes in this thesaurus with many children tend to acquire more children over time. Chien *et al.* [7] examined PubMed articles in medicine and health using the x-index. Ilgisonis *et al.* [22] analyzed descriptors to investigate trends in general and personalized medicine. Several studies examined descriptor co-occurrence, particularly in relation to new descriptors [1,24]. Kastrin and Hristovski [24] built a network from descriptor co-occurrences to investigate indexing trends. Anastasios *et al.* [1] demonstrated that approximately one-quarter of new descriptors in MeSH represent emerging concepts. Ilgisonis *et al.* [22] further focused on related descriptors, defined as descriptors that co-occur with annotated descriptors. Despite these contributions, the analysis of instances—such as the occurrence of descriptors in PubMed articles—remains largely unexplored for quantifying MeSH concepts over time.

This study addresses this gap by proposing an approach to measure the relevance of MeSH concepts through their instantiation in biomedical literature. The approach integrates three main factors: the usage of concepts in PubMed

articles, the hierarchical structure of MeSH, and the citation network of PubMed articles, analyzed through information-theoretic and network-based methods. We define four dimensions of relevance: informativeness, usefulness, disruptiveness, and influence. To quantify informativeness, we apply entropy from information theory [16,40]. To measure usefulness, we use category utility supported by the hierarchical organization of MeSH [11]. To assess disruptiveness, we consider whether concepts annotate highly cited but consolidated technologies, which may not contribute novelty to the field [13,15,43]. To evaluate influence, we apply centrality metrics from graph theory [33], which capture the importance of nodes in networks [19]. We apply these methods to the MEDLINE database and analyze MeSH annotations over time. We generate monthly data by mapping references to MeSH concepts and computing each metric for each concept. We propagate this data through the MeSH hierarchy and combine the multidimensional scores using reciprocal rank fusion (RRF) to estimate the relative relevance of MeSH concepts.

This paper presents three main contributions:

1. We propose an algorithm to compute the relevance of MeSH concepts by propagating information from a network of instances to the KOS structure. The network of MeSH instances is based on the PubMed citation network, while the knowledge structure relies on the thesaurus hierarchy.
2. We assess the relevance of MeSH concepts quantified by multidimensional metrics—informativeness, usefulness, disruptiveness, and influence—and show how they vary over time for different hierarchical levels.
3. We provide a quantitative evaluation of the approach in two use cases: MeSH terminology evolution and analysis of retracted papers. We compare the relevance of MeSH terms in a test group (concept changes and retracted papers) versus a control group (no evolution and no retraction, respectively).

2 Materials and Methods

We explore the evolution of MeSH by analyzing concept instantiations across multiple years in PubMed. For this purpose, we use the PubMed/MEDLINE corpus, which contains articles manually annotated with MeSH descriptors. In this section, we present the methodology used to create our dataset. We also describe the metrics used to quantify relevance in a KOS. Next, we discuss strategies to propagate these metrics through the hierarchy and to compute a metrics fusion. Finally, we outline the evaluation criteria applied in this work.

2.1 Dataset and Pre-processing

We conducted our study using the 2022 collection of PubMed[1], which comprises 33,403,054 articles published up to December 2021. PubMed indexes new biomedical and life science articles as soon as they are published. Articles in

[1] https://lhncbc.nlm.nih.gov/ii/information/MBR/Baselines/2022.html.

PubMed are annotated by domain experts using MeSH descriptors, which are organized hierarchically into 16 categories denoted by single letters.

For this collection, we built citation networks using PubMed article citations and the iCite[2] dataset [21] as a more comprehensive source of citation data. We then created cumulative monthly networks, i.e., incorporating nodes from previous months, from January 2014 through December 2021. To build the corresponding MeSH instance hierarchy, we used all the MeSH descriptors assigned to articles in a particular month and mapped them to their respective MeSH tree codes. The first monthly citation network (January 2014) contained 18,728,651 nodes and 381,296,042 edges, while the last one (December 2021) contained 27,728,971 nodes and 673,415,629 edges. Because computing certain metrics was prohibitively time-consuming, we randomly sampled 10% of the nodes from each original monthly citation network. Edges were retained only when both nodes appeared in the sampled network. Thus, with this approach, we preserve the structure similar to the full citation network. This resulted in the first sampled citation network (January 2014) having 1,872,866 nodes and 3,807,856 edges, while the last sampled network (December 2021) had 2,772,898 nodes and 6,745,896 edges.

2.2 MeSH Relevance Quantification

Analogous to the definition of *relevance* in information science and information retrieval [30], we define the relevance of a concept as the extent to which a concept or a set of concepts meets the representational needs of a dataset [39]. Following the definition of relevance provided by Saracevic [38], in our study, we consider multiple aspects to quantify the relevance of MeSH concepts, namely, informativeness, usefulness, disruptiveness, and influence (Fig. 1, Metrics). We provide the code to compute these aspects[3] as well as the computed metrics[4]. Each PubMed article is annotated with MeSH descriptors, which map to one or more tree nodes. In the following, we describe these metrics and explain how the relevance of a MeSH concept is quantified according to them.

Informativeness. To quantify the relevance of a concept in terms of informativeness, we use entropy, introduced by Shannon [16,40]. Entropy quantifies the level of uncertainty of a variable. In our case, if a concept is frequently used to annotate many papers, it is less informative, since numerous other articles also contain this annotation. For example, the concept "Humans" (D006801) is used to annotate around 70% of PubMed articles. Therefore, this concept contributes little information when it appears in an article and has a low entropy value. In contrast, the concept "Proteins" (D011506) is used to annotate only 2% of PubMed articles, which makes it considerably more informative.

[2] https://nih.figshare.com/articles/dataset/iCite_Database_Snapshot_2022-04/19763386.
[3] https://github.com/jcoparaz/computation_metrics.
[4] https://zenodo.org/records/13635979.

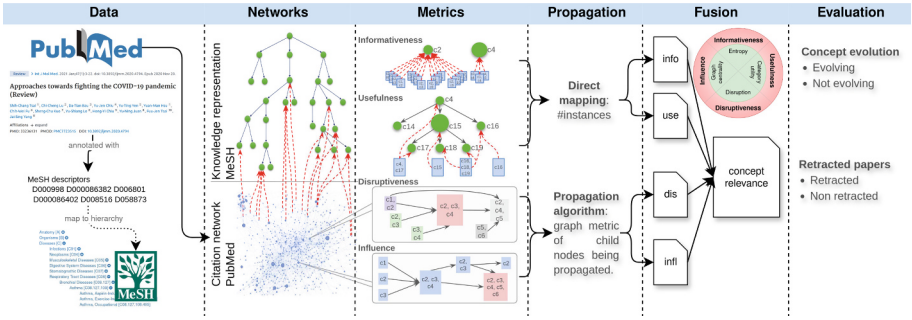


Fig. 1. Pipeline of concept relevance quantification in MeSH.

As shown in Eq. 1, we follow Shannon's entropy definition $H(X)$, where $p(x)$ represents the probability of an event's occurrence and x refers to all values that a variable X can take:

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \tag{1}$$

In our study, the instantiation of a MeSH concept represents an event, that is, articles mapped to nodes in the MeSH hierarchy, representing the concept's usage (dotted red arrows in Fig. 1, Networks). The probability of an event $p(x)$ is calculated as the probability that a concept x is mapped within all mappings in its hierarchical level (its distance to the root node). Mappings at lower levels are propagated to their parents when calculating the latter's probability.

Usefulness. To quantify relevance in terms of usefulness, we use the category utility measure to identify basic-level categories in a taxonomy [11]. Basic-level categories are preferred in both natural and artificial hierarchies, and they carry the most information about the category's attributes [11,37] (e.g., when shown an image of a specific oncological exam, most professionals identify it as cancer rather than as a general disease or a colon cancer). A node in a hierarchy is useful to the extent that it improves the ability to predict the features of instances in that category. As shown in Fig. 1, in this study, the mapping of articles to the hierarchy characterizes concepts. Therefore, a concept within a densely mapped branch in the hierarchy is considered more useful (e.g., as node *c15* in the example).

Category utility combines the frequency of a category's occurrence with the probability of its features [11]. Let us define a category c as a finite set of instances: $c = \{o_1, o_2, ..., o_n\}$, where each instance o_i can be described by a finite set of discrete features $F = \{f_1, f_2, ..., f_m\}$. The category utility (CU) of category c with a set of features F can then be calculated as in Eq. 2, where $p(c)$ is the probability of a category (i.e., a MeSH concept in our case), $p(f_k)$ is the probability of feature f_k and $p(f_k|c)$ is the conditional probability of feature f_k given category c:

$$CU(c, F) = p(c) \sum_{k=1}^{m} [p(f_k|c)^2 - p(f_k)^2]. \quad (2)$$

To calculate category utility for the nodes in the MeSH hierarchy, we use mapping to articles a_i as the set of features $F = \{a_1, a_2, ..., a_m\}$, where each feature a_i has a binary value (annotated or not with a category). To compute the probabilities, we created a binary matrix M_{ij} of n tree nodes and m features as follows:

$$M_{ij} = \begin{cases} 1, & \text{tree node } i \text{ is mapped by article } j, \\ 0, & \text{otherwise.} \end{cases}$$

Disruptiveness. To measure the relevance of a concept in terms of the aspect of disruptiveness, we use the disruption index [13]. This index was originally designed to quantify the extent to which an invention (a concept, in our case) consolidates or destabilizes the subsequent use of the components (i.e., referenced concepts) on which it builds [13]. To compute the disruption index D for a focal node, we consider the citation network as a tripartite graph $G = (V_1, V_2, V_3, E)$, where the set of nodes V_1, V_2, V_3 represents the articles, which are connected by directed edges E, representing their citations [19,33]. Following Figueiredo and Andrade's definition, given a focal node a (light red in Fig. 1, Metrics, Disruptiveness), corresponding to an article for which we want to compute the level of disruptiveness, there are j articles that reference a and at least one of predecessor (purple box), i articles that reference a but none of its predecessors (green box), and k articles that do not reference a but reference at least one predecessor (dark gray box), disruption D for a is defined by Eq. 3. D can take values in the range [-1,1].

$$D = \frac{n_i - n_j}{n_i + n_j + n_k}, \quad (3)$$

where n_i, n_j, and n_k represent the number of i, j, and k articles, respectively. The disruption score D of an article is then propagated to the concepts mapped to it ($c2$, $c3$, and $c4$ in the example). Finally, the disruption index D_c for a concept c is the sum over the disruption scores for all articles $a_1, \cdots, a_n$ which are mapped to c, divided by the number of articles in the citation network, i.e., $D_c = \sum_{i=1}^{n} D_i/m$, where D_i is the disruption index of the ith article mapped to c and m is the number of articles in the citation network. To compute D, we used Figueiredo and Andrade's disruption implementation [13].

Influence. In graph theory, the influence of a node is quantified through graph centrality, based on the importance of neighboring nodes. We use this concept to measure the influence aspect of a MeSH concept. Rather than relying on the MeSH hierarchy, which captures only ontological relations, we use instances (i.e., mapped articles) to identify the influence of MeSH concepts. Similar to the

disruptiveness aspect, we consider the citation network to compute centrality. In Fig. 1, we show a small citation network example to demonstrate how nodes (articles) influence their neighbors. An article annotated with one or more concepts, denoted by c_1, ..., c_n, gains influence from its incoming articles (e.g., light red box).

In our work, we used the PageRank algorithm to compute the centrality of a node [3,33], which uses the importance of the node's neighbors to compute the node's importance. The PageRank of node i in a directed graph G is defined as x_i in Eq. 4:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta, \tag{4}$$

where the parameter α regulates the contribution between the eigenvector and the constant term β. The term β is used to initialize nodes with zero out-degree. The adjacency matrix of G is denoted by A_{ij}, which captures the edges between pairs of nodes (citations in our case). The out-degree of node j is represented by k_j^{out}. Similar to disruptiveness, the centrality x_i of an article i is propagated to the concepts it is mapped to, and the overall centrality x_c of a concept is computed by summing over all the centralities of articles that concept c is mapped to divided by the number of articles in the citation network, i.e., $x_c = \sum_{i=1}^{n} x_i/m$, where x_i is the centrality score of the ith article mapped to c and m is the number of articles in the network. We used the PageRank implementation from the NetworKit library[5] to calculate the graph centrality, where $\alpha = 0.85$ and $\beta = 1 - \alpha$.

2.3 Propagation of Data Instances to the Knowledge Representation

We encoded the usage of MeSH instances and propagated them to the knowledge representation for each aspect. Given the "*broader than*" relationship between MeSH concepts, we consider two methods for propagating information to higher-level nodes: direct mapping or a cumulative weighted propagation algorithm. For the information-theoretical measures (entropy and category utility), which are computationally efficient, we directly propagate instance mappings to their parent concepts using the tree nodes associated with each descriptor. Then, we compute the relevance scores for each node in the hierarchy.

We approach the network analysis metrics (disruption and graph centrality) differently due to their high computational complexity. Instead of combining the mappings into higher-level nodes, we aggregate the scores of lower-level nodes into those of their parents. The assumption is that if a node is disruptive or influential, so too will its parent. As described in Algorithm 1, the propagation method requires the MeSH hierarchy H and the network-based relevance scores computed using the citation network (variable tnm) as input. In line 2, we obtain tree nodes per level of specificity using the hierarchy H; for instance, tree nodes C and D are in level 1, while tree nodes C01 and D01 are in level 2. Then,

[5] https://networkit.github.io/.

we iterate over the hierarchy from bottom to top, using a reversed level list (variable *llr*, line 3) to propagate the relevance scores of the lower-level concepts to the higher levels. First, at each level, we get a set of unique parents (lines 5–9). Then, we retrieve their children (line 11) and, utilizing their initial measurements (lines 14–18), compute the weighted parent relevance, variable *gmp*, based on the children's relevance and the number of nodes in their level (lines 19 and 21). Finally, the parent value is calculated by taking into account the direct mapping to its node in addition to the disruption of child nodes being propagated $d_{n_i} = direct_d_{n_i} + indirect_d_{n_i}$ (lines 22 and 23).

Algorithm 1 Propagation of graph-based metrics to the MeSH hierarchy

```
Require: H (hierarchy), tnm (graph metric values initialized from descriptors)
Ensure: gmh (graph metric of tree nodes in the hierarchy)
1: gmh ← {}
2: dpt ← get_nodes_in_levels_specificity(H)
3: llr ← sorted list of keys in dpt, reversed
4: for each current_level in llr do
5:     unique_parents ← an empty set
6:     for each treenode in dpt[current_level] do
7:         parent ← get_parent(H, treenode)
8:         add parent to unique_parents
9:     end for
10:    for each parent in unique_parents do
11:        children ← get_children(H, parent)
12:        gmp ← 0
13:        for each child in children do
14:            if child ∈ tnm then
15:                if child has no children then
16:                    gmh[child] ← tnm[child]
17:                end if
18:            end if
19:            gmp ← gmp + gmh[child]
20:        end for
21:        gmp ← gmp/len(dpt[current_level])
22:        dgm ← tnm[parent] if parent ∈ tnm else 0
23:        gmh[parent] ← gmp + dgm
24:    end for
25: end for
26: return gmh
```

2.4 Consolidating the Relevance Ranking of MeSH Codes

Each of the metrics described previously measures different relevance aspects of a MeSH concept according to its data instances. To have a unified relevance score, we propose fusing the multidimensional aspects into a single ranking. Inspired by ranking fusion methods in information retrieval, we consider that each concept $d \in D$, where D is the set of MeSH codes, is ranked according to a relevance dimension $r \in R$, where R is the set of four aspects considered in this work: informativeness, usefulness, disruptiveness, and influence.

Information retrieval research proposes various techniques for merging rankings, such as CombSUM, CombMNZ, CombANZ, rank-biased precision, and reciprocal rank fusion (RRF) [10]. Due to its simplicity, effectiveness, and robustness [10,41], we use RRF to compute the overall relevance of a MeSH concept based on its instantiations. RRF combines the outcomes of different information retrieval methods with varying relevance scores into a single result set using the rank information. The mathematical definition of RRF is provided in Eq. 5, where, in our case, a concept d belongs to the set of MeSH codes D, and $r(d)$ represents the rank of a code d according to a metric r from the set of metrics R. The constant k minimizes the impact of outliers. We set k to 60, as defined in the original RRF paper [10,41].

$$\mathrm{RRF}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}, \tag{5}$$

2.5 Evaluation Criteria

We qualitatively analyze the trends of MeSH concepts according to the different relevance aspects for various levels of the hierarchy and across MeSH versions from 2014 to 2021. To quantitatively evaluate the effectiveness of our approach, we conducted two experiments. First, we assess the relevance of concepts that changed against those that remained unchanged between MeSH versions. Second, we compare the relevance of concepts between retracted and non-retracted papers in PubMed. We used the Mann-Whitney test to compare relevance changes; *p-values* below 0.05 are considered statistically significant.

3 Results

This section presents results assessing the relevance of MeSH concepts using a multi-aspect model. It includes an analysis of the evolution of the MeSH terminology, highlighting key concepts at different hierarchy levels. Two use cases are explored: i) comparing the relevance of concepts that changed versus those that remained unchanged, and ii) comparing concepts used to annotate retracted papers versus those used for non-retracted papers.

3.1 MeSH Concept Relevance Trends

We calculated the relevance of MeSH concepts from 2014 to 2021 using the PubMed 2022 snapshot annotated with the MeSH 2022 version. Relevance scores for the different aspects—disruptiveness, influence, informativeness, and usefulness—as well as for the overall concept relevance were computed monthly. Figure 2a depicts the cumulative trends, i.e., the sum of the first level of the MeSH hierarchy (A, ..., Z). Disruptiveness, influence, and usefulness decrease over time ($p < 0.001$). Park and colleagues also corroborate the decreasing trend of disruptiveness, although their analysis was based on scientific articles, in contrast to ours, which is based on MeSH concepts [36]. On the other hand, informativeness, which measures the degree of uncertainty, tends to increase over time ($p < 0.001$). Therefore, the degree of uncertainty also increases as more recent knowledge is added to PubMed.

Figure 2b shows the trend of the concept relevance at the first level in MeSH. All of the categories show a decreasing trend ($p < 0.001$). The most useful concepts (Fig. 2b - bottom right) in the first level are D (Chemicals and Drugs), followed by N (Health Care) and C (Diseases), indicating they encode the basic-level categories of PubMed. In influence, all categories show a decreasing trend ($p < 0.001$). Concept C is the most influential (Fig. 2b - top right), indicating that 'Diseases' is a central concept in PubMed. It is followed by the concepts D

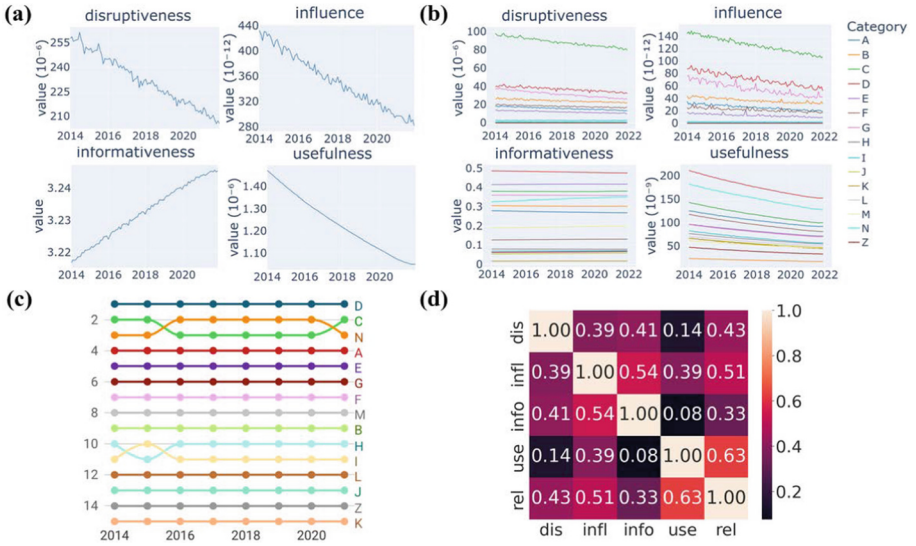


Fig. 2. Trend of concept relevance in MeSH at level 1 and correlation of descriptors. a) The sum of MeSH categories. b) Trend of MeSH categories. c) Ranking of concepts at level 1. d) Correlation of aspects. Column names are dis: disruptiveness (disruption), infl: influence (centrality), info: informativeness (entropy), use: usefulness (category utility), rel: concept relevance (RRF fusion).

and G (Phenomena and Processes), which include chemical, and physiological phenomena descriptors.

Trends in disruptiveness are mostly decreasing ($p < 0.001$), though categories M and H are not statistically significant. J shows an increasing trend ($p < 0.001$). On the other hand, K shows a growing trend, while I shows no significant trend. Concept C is also the most disruptive in the first level of the MeSH hierarchy (Fig. 2b - top left), being more than twice as disruptive as the second-ranked concept (D). Trends of informativeness are primarily increasing with a $p < 0.001$. However, categories H, G, B, A, and D present a decreasing trend ($p < 0.001$). Category C shows no trend (not statistically significant). Concept D has the highest informativeness score (Fig. 2b - bottom left), followed by concept E (Analytical, Diagnostic and Therapeutic Techniques, and Equipment). However, its score gradually decreases, with a standard deviation of 0.003377. The trends of this aspect remain relatively constant over time, largely due to the narrow range of values across MeSH categories.

Figure 2c shows the ranking of concepts in level 1. Most concepts retain their positions over time, except for Diseases (C), Health Care (N), Disciplines and Occupations (H), and I, which show some variation. Chemicals and Drugs (D) consistently ranks higher than other MeSH categories. The second position in this ranking alternates between 'Diseases' and 'Health Care', with the latter prevailing in five of eight years.

Figure 2d shows the correlation between relevance aspects for each descriptor using the monthly values in releases. The overall concept relevance, as a fusion of disruptiveness, influence, informativeness, and usefulness, shows a moderate correlation. Usefulness has a low correlation with disruptiveness and informativeness and a moderate correlation with influence. Influence and informativeness exhibit a moderate correlation.

3.2 Top Concepts in the MeSH Terminology

Note that some rank shifts reflect relative movements; small score differences can translate into large changes in rank due to the discretization of values. Figure 3a displays the top 10 concepts according to the concept relevance, at level 2 (left side), level 3 (center), and level 4 (right). Half of the top concepts maintain their ranking over time. Concept G02 (Chemical Phenomena) shows variation among several positions, starting in fourth place in 2014 and ending in eighth in 2021. Meanwhile, C12 (Urogenital Diseases), D01 (Inorganic Chemicals), D02 (Organic Chemicals), and C01 (Infections) change by one position over time.

In level 3, four concepts maintain a stable ranking, with B01.050 (Animals) leading. Concepts E05.481 (In Vitro Techniques), A03.620 (Liver), E01.789 (Prognosis), and N06.850 (Public Health) move among multiple positions, with 'Prognosis' tending to climb positions over time. In level 4, only the top three concepts, M01.060.116 (Adult), M01.060.406 (Child), and M01.060.057 (Adolescent), remain stable over time. These concepts continue to generate high research interest. The remaining concepts mainly decrease by one position in the ranking.

We computed the rank trend of concepts by adding the differences between consecutive rank positions of a concept and dividing them by the number of differences, i.e., the number of years minus 1. Thus, we identified the concepts with the highest and lowest slopes. The top 10 upward concepts are shown in Fig. 3b, which seem to gain positions over time at levels 2, 3, and 4. The behavior of the rank trend of concepts in these levels is similar. The concepts with the highest slope in level 2 are D06 (Hormones, Hormone Substitutes, and Hormone Antagonists), I01 (Social Sciences), D05 (Macromolecular Substances), and C24 (Occupational Diseases). The concept with the highest slope in level 3 is D01.930 (Thorium Compounds), and in level 4 is D03.383.517 (Oxathiins). The downward-trending concepts are shown in Fig. 3c, which appear to lose positions over time, in contrast to the upward concepts. The concepts with the lowest rank trend slope are G16 (Biological Phenomena) and I02 (Education), both in level 2. In level 3, the concept M01.928 (Vegetarians) has the lowest trend slope, while in level 4, the concept E05.200.312 (COVID-19 Testing) has the lowest slope.

The bottom concepts are shown in Fig. 3d. Concepts in level 2 remain stable, while in levels 3 and 4, concepts change positions constantly. On the other hand, bottom concepts in levels 3 and 4 vary more widely, often by hundreds of ranks, as in these levels, there are more concepts. Most of the changes occur towards the end of the studied period. The concepts J02.750 ('Sustenance,' in level 3) and B03.280.622 ('Pannus,' in level 4) seem to gain positions over time. In Fig. 3d, right panel, we are only showing the plot since 2017 for clarity, as previous years

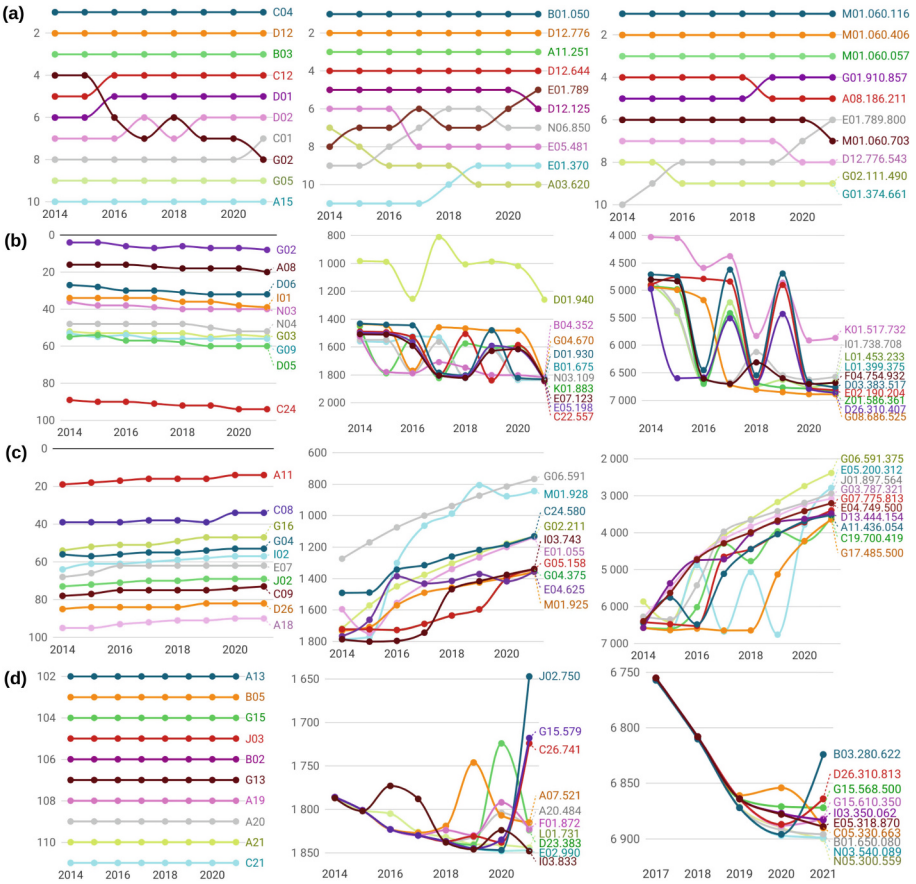


Fig. 3. Top/bottom concepts over time per level 2 (left), 3 (center), and 4 (right). The left side of each figure shows the rank position of the concept. a) The top 10 concepts with the highest average rank. b) Concepts with the highest rank trend slope. c) Concepts with the lowest rank trend slope. d) The bottom 10 concepts with the lowest average rank.

follow the trend of 2017–2018. The MeSH headings of the concepts in Fig. 3 can be found in the MeSH tree [34].

3.3 Quantitative Assessment

In this section, we describe the quantitative assessment of the multiple relevance aspects through the concept evolution task and retracted papers. We used an ontology evolution dataset for the concept evolution task and identified retracted papers in PubMed. We first describe the datasets, then detail each assessment and present results. We chose the Mann-Whitney test as our data does not follow a normal distribution [29].

Datasets for Quantitative Assessment.

Ontology Evolution Dataset. We created our dataset for concept evolution as previous studies still need to release their datasets. This dataset contains releases where MeSH descriptors are annotated with the type of change and data-driven temporal features. We calculated these features from the Data-properties dataset (explained in Sect. 2.1).

A MeSH descriptor (or concept) is considered to have evolved if it underwent a change in description (change of description), gained new children (extension), changed location (move), or was removed (removal) [6]. We use the ontology versions of MeSH available in BioPortal, which have two versions per year, e.g., 2014AA and 2014AB releases. When we downloaded the MeSH releases[6], releases were available from 2013 to 2021.

We used COntoDiff [20] to get the difference between two consecutive MeSH releases and found no changes in 2013. Thus, we excluded this year. We identified which concepts changed in each release and their type of change (description, extension, move, and removal). We used data until 2021AA, as we used 2021AA and 2021AB to identify the concepts that changed in 2021AA. Evolving concepts represent about 1% of concepts per release, and releases AB have no or very few changes. Over releases, there are $28,900$ concepts on average, the average relevance for evolving concepts is $2.09E-03$, while for concepts that did not evolve is $8.46E-04$.

Retracted Papers in PubMed. We identified the retracted papers in the entire collection of PubMed 2022. Then, we identified the retracted papers in the monthly citation networks. On average, 800 papers were found to be retracted per month. We then retrieved the MeSH annotations of each article, along with the tree nodes associated with each MeSH descriptor. We obtained a value per metric for each article using these tree nodes by summing their metrics' values. Since our citation networks are based on a random sample, a paper may appear in multiple months. We therefore averaged each metric per paper per year. We used these values to perform the statistical test.

Assessment in Concept Evolution. We focused on the concept evolution task to investigate whether the approach proposed in this work could contribute to biomedical terminology evolution. We performed a statistical test of significance using the Mann-Whitney test to determine if there is a significant difference between the relevance of descriptors that changed and descriptors that did not change among releases. For the concept relevance (RRF), we found statistical significant differences ($p < 0.05$) whenever evolving descriptors were present.

Figure 4a shows the test results for each relevance aspect in every release. Releases 2016AB, 2017AB, 2019AB, and 2020AB do not contain evolving descriptors. Consequently, statistical tests could not be applied. Five out of eleven releases exhibit statistically significant differences in usefulness, while in

[6] Downloaded on April 12, 2022.

influence and disruptiveness, eight show statistically significant differences. All releases show statistically significant differences in informativeness, but only one release in concept relevance does not demonstrate a statistical difference.

(a)

Release	Concept relevance
2014AA	$p < 0.001$
2014AB	$p < 0.001$
2015AA	$p < 0.001$
2015AB	$p = 0.29$
2016AA	$p < 0.001$
2016AB	–
2017AA	$p < 0.001$
2017AB	–
2018AA	$p < 0.001$
2018AB	$p = 0.04$
2019AA	$p < 0.001$
2019AB	–
2020AA	$p < 0.001$
2020AB	–
2021AA	$p < 0.001$

(b)

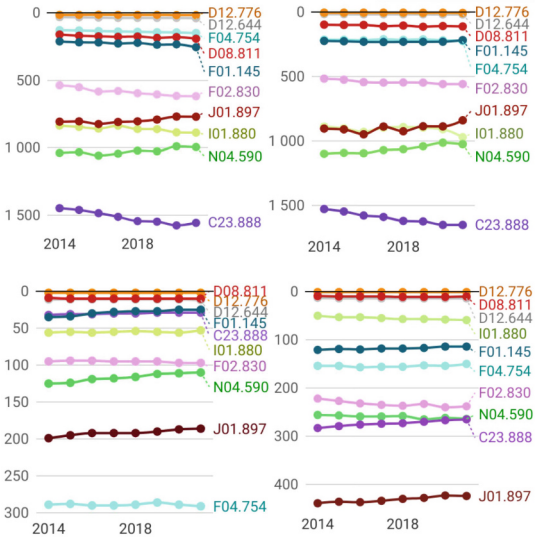


Fig. 4. Evaluation of the multiple aspects in concept evolution. a) Results of the statistical significance test for each release. Light gray-coded cells correspond to $p < 0.05$; otherwise, it is dark gray. b) The top evolving concepts according to each aspect are (starting top left) disruptiveness, influence, informativeness, and usefulness. (Color figure online)

Figure 4b shows the concepts that mostly evolve according to each aspect. Some concepts change their position in disruptiveness and influence. Concepts in informativeness and usefulness remain mostly constant. The concepts D12.776 (proteins), D12.644 (peptides), and D08.811 (enzymes) are among the top three aspects. The MeSH headings of the remaining concepts can be found in the MeSH tree [34].

Assessment in Retracted Papers. We used the multiple aspects metrics in this study to determine whether there is a significant difference between retracted and non-retracted papers in PubMed. The results (Fig. 5a) show significant differences between retracted and non-retracted papers in concept relevance, using the proposed aspects as the comparison unit, with a $p < 0.05$. Furthermore, when we performed the test by aspect, we also found a statistical difference in usefulness. In influence, there is a significant difference in five out of eight years, while in disruptiveness and informativeness, seven years show differences.

Figure 5b shows the top concepts in retracted papers according to the rank relevance at level 3. Concepts B01.050 (Animals), A11.251 (Cells, Cultured), and E01.789 (Prognosis) lead the ranking. Half of the concepts come from branch G, and three belong to branch G04 (Cell Physiological Phenomena). The MeSH headings of the remaining concepts can be found in the MeSH tree [34].

4 Discussion

Previous research on MeSH has produced useful insights, though with limitations in capturing the multidimensional nature of concept relevance. Structural approaches of hierarchical growth patterns provide insight into organizational principles but do not reflect actual scientific impact or usage patterns [2]. In bibliometric and co-occurrence methods, researchers have attempted to bridge the gap between structure and usage [7,22,24]. However, these approaches treat concepts as isolated units without considering citation relationships or hierarchical propagation.

(a)

Year	Disruptiveness	Influence	Informativeness	Usefulness	Concept relevance
2014	$p = 0.02$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2015	$p = 0.68$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2016	$p = 0.03$	$p = 0.002$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2017	$p < 0.001$	$p = 0.56$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2018	$p < 0.001$	$p = 0.69$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2019	$p < 0.001$	$p = 0.38$	$p < 0.001$	$p < 0.001$	$p < 0.001$
2020	$p < 0.001$	$p < 0.001$	$p = 0.15$	$p < 0.001$	$p < 0.001$
2021	$p < 0.001$	$p < 0.001$	$p = 0.05$	$p < 0.001$	$p < 0.001$

(b)

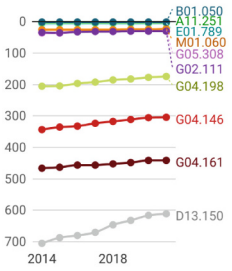


Fig. 5. Evaluation of the aspects in retracted papers. a) Results of the statistical significance test. Light gray-coded cells correspond to $p < 0.05$; otherwise, it is dark gray. b) Top-10 concepts in retracted papers according to the concept relevance at level 3. (Color figure online)

Insights from the knowledge graph community offer complementary tools for ranking the importance of classes, i.e., concepts [12]. These approaches often overlook the rich citation dynamics available in bibliographic resources such as PubMed. Furthermore, studies on scientific disruptiveness indicate that citation patterns can reveal whether a topic displaces prior work or builds on it, insights that raw term counts fail to capture [43]. Finally, practical work on annotation and thesaurus maintenance highlights the need to prioritize which concepts to review or update in real-world applications [4].

A critical gap exists in instance-based analysis, where, despite millions of PubMed articles providing rich instantiation data for MeSH concepts, this resource has remained largely unexploited. Unlike previous work that relies on

related or co-occurring descriptors, our approach leverages actual MeSH annotations as concept instantiations, providing a measurement of how concepts are employed in scientific discourse. By integrating information theory with network analysis and hierarchical propagation, our multidimensional framework addresses the limitations of single-perspective approaches, captures temporal evolution, and provides systematic validation through MeSH terminology changes and retracted articles.

Leveraging concepts from information theory (i.e., entropy and category utility) and network analysis (i.e., disruptiveness and graph centrality), we investigated different aspects of the relevance of MeSH concepts. Data instances are propagated through direct mapping for information-theoretical measures and Algorithm 1 for network analysis metrics. Direct mapping considers instances when computing these measures. To summarize these four metrics in an overall concept relevance ranking, we applied the RRF method. In our experiments, we analysed the relevance of MeSH concepts to the PubMed corpus, using the 2022 snapshot annotated with MeSH 2022. Using monthly data from 2014 to 2021, we computed the relevance aspects and showed how they vary over time. We applied the methodology to two use-cases—concept evolution and retracted articles—and analyzed the MeSH concept relevance in these scenarios.

Overall, there is an increasing trend for informativeness and a decreasing trend for usefulness, influence, and disruptiveness. The latter is similar to a recent finding from a study of Park *et al.*, which shows that disruptiveness in life and biomedicine sciences is decreasing [36]. At the high level of the MeSH categories, our results show that categories C (*Diseases*) and D (*Chemicals and Drugs*) are the most relevant for the PubMed collection. While this result is expected, given the importance of such concepts in life and biomedical sciences, our method can quantify their relevance according to different aspects, as well as an aggregated relevance rank. In particular, category D ranks first in terms of informativeness and usefulness, while category C leads in influence and disruptiveness. The correlation analysis of descriptors revealed a weak correlation among informativeness, usefulness, and disruptiveness, demonstrating that indeed they capture different aspects of concept relevance.

Analysis of the top 10 relevant concepts shows that higher levels are more stable than lower ones due to the propagation of lower-level relevance scores; for instance, level 1 is more stable than level 2. At level 2 (three-digit codes), concepts undergo notable ranking shifts, with Chemical Phenomena (G02) falling by four positions between 2014 and 2021. Levels 3 (six digits) and 4 (nine digits) are even more dynamic: Prognosis (E01.789) rises three positions, while Liver (A03.620) drops three in the same period, and most level 4 concepts lose one position. Within categories C, D, and E, level 2 remains largely stable, but at level 3, Cardiovascular Diseases (C14) and Investigative Techniques (E05) show the greatest changes, possibly reflecting evolving scientific knowledge. Overall, lower-ranked concepts (bottom 10) shift more than higher-ranked ones, except at level 1.

We further evaluated the application of our proposed approach using two analysis use-cases: concept evolution and retracted papers. For the former, we utilized an ontology evolution dataset, containing concept changes from consecutive MeSH versions. Interestingly, the Mann-Whitney test confirmed statistically significant differences between the relevance rank of MeSH concepts that evolve versus those that remain stable. This result suggests that the relevance of a concept could be used to identify candidates for update in KOS maintenance. Additionally, we show that retracted and non-retracted papers can be differentiated using our approach even though the monthly proportion of retracted papers was approximately 0.0002%. In particular, concepts *Animals* (B01.050), *Cells, Cultured* (A11.251), *Prognosis* (E01.789), *Age Groups* (M01.060), *Gene Expression Regulation* (G05.308), and *Biochemical Phenomena* (G02.111) showed consistently high relevance to the retracted corpus. While the former three concepts also appear in the overall top 10 relevance list, the latter three concepts are particular to retracted papers, suggesting potential life and biomedical research areas with issues.

This study has several limitations. First, extending this work to other domains requires a dataset of instances annotated with a KOS, such as PubMed with MeSH. Alternatively, information extraction methods could be used to automatically annotate instances or corpus [8,9,31]. Second, all metrics were computed using a sample from the citation network due to the algorithmic complexity of some metrics, which made full computation prohibitive. Finally, metrics are somewhat correlated. Within a short period (e.g., one release to the next), our method is not always sensitive enough to identify all changes.

5 Conclusions

In this paper, we present a methodology based on information theory and network analysis to compute the relevance of MeSH concepts for the biomedical corpus indexed by PubMed. By considering multiple relevance aspects of a concept in a data-driven analysis, our approach provides new insights into the evolution of MeSH concept relevance within PubMed.

We analyzed the dynamics of MeSH over time and showed the potential of our approach in tasks such as tracking concept evolution and identifying retracted papers. We found statistically significant differences in relevance values across these tasks. Future work should extend this methodology to other domains and address the computational complexities associated with some of the metrics.

Acknowledgments. Part of this work has been funded by the CINECA project (H2020 No 825775) and Innosuisse project funding number 46966.1 IP-ICT.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anastasios, N., Anastasia, K., Grigorios, T., Georgios, P.: What is all this new MeSH about? Int. J. Digit. Libr. **22**(4), 319–337 (2021)
2. Balogh, S.G., Zagyva, D., Pollner, P., Palla, G.: Time evolution of the hierarchical networks between pubmed MeSH terms. PLoS ONE **14**(8), e0220648 (2019)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
4. Cardoso, S.D.: MAISA - Maintenance of semantic annotations. Theses, Université Paris Saclay (COmUE) (2018). https://theses.hal.science/tel-02288589
5. Cardoso, S.D., Da Silveira, M., Pruski, C.: Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies. Knowl.-Based Syst. **194**, 105508 (2020)
6. Cardoso, S.D., Pruski, C., Da Silveira, M.: Supporting biomedical ontology evolution by identifying outdated concepts and the required type of change. J. Biomed. Inform. **87**, 1–11 (2018)
7. Chien, T.W., Wu, H.M., Wang, H.Y., Chou, W.: The most cited MeSH terms and authors who published papers in Pubmed Central on the topic of medicine and health using bibliometric analyses. Asian J. Med. Health **14**(4), 1–9 (2019). https://doi.org/10.9734/ajmah/2019/v14i430105
8. Copara, J., Knafou, J., Naderi, N., Moro, C., Ruch, P., Teodoro, D.: Contextualized french language models for biomedical named entity recognition. In: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, pp. 36–48 (2020)
9. Copara, J., Naderi, N., Knafou, J., Ruch, P., Teodoro, D.: Named entity recognition in chemical patents using ensemble of contextual language models. In: Proceedings of the CLEF 2020 Conference. 22-25 September 2020 (2020)
10. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 758–759 (2009)
11. Corter, J.E., Gluck, M.A.: Explaining basic categories: feature predictability and information. Psychol. Bull. **111**(2), 291 (1992)
12. Fernández-Álvarez, D., Frey, J., Labra Gayo, J.E., Gayo-Avello, D., Hellmann, S.: Approaches to measure class importance in knowledge graphs. PLoS ONE **16**(6), e0252862 (2021)
13. Figueiredo, F., Andrade, N.: Quantifying disruptive influence in the AllMusic guide. In: Proceedings of the 20th International Society for Music Information Retrieval Conference, pp. 832–838. ISMIR (2019). https://doi.org/10.5281/zenodo.3527940
14. Fortunato, S., et al.: Science of science. Science **359**(6379), eaao0185 (2018)
15. Funk, R.J., Owen-Smith, J.: A dynamic network measure of technological change. Manage. Sci. **63**(3), 791–817 (2017). https://doi.org/10.1287/mnsc.2015.2366
16. Ghahramani, Z.: Information Theory. John Wiley & Sons, Ltd (2006). https://doi.org/10.1002/0470018860.s00643
17. Gobeill, J., Pasche, E., Teodoro, D., Veuthey, A.L., Lovis, C., Ruch, P.: Question answering for biology and medicine. In: 2009 9th International Conference on Information Technology and Applications in Biomedicine, pp. 1–5. IEEE (2009)

18. Gobeill, J., Teodoro, D., Pasche, E., Ruch, P.: Taking benefit of query and document expansion using MeSH descriptors in medical imageclef 2009. In: Conference and Labs of the Evaluation Forum (2009)
19. Hamilton, W.L.: Graph representation learning. Morgan & Claypool Publishers (2020)
20. Hartung, M., Groß, A., Rahm, E.: COnto-Diff: generation of complex evolution mappings for life science ontologies. J. Biomed. Inform. **46**(1), 15–32 (2013)
21. iCite, Hutchins, B.I., Santangelo, G., Haque, E.: iCite Database Snapshot 2022-04 (2022). https://doi.org/10.35092/yhjc.19763386.v1
22. Ilgisonis, E.V., Pyatnitskiy, M.A., Tarbeeva, S.N., Aldushin, A.A., Ponomarenko, E.A.: How to catch trends using MeSH terms analysis? Scientometrics **127**(4), 1953–1967 (2022)
23. Ivanović, M., Budimac, Z.: An overview of ontologies and data resources in medical domains. Expert Syst. Appl. **41**(11), 5158–5166 (2014)
24. Kastrin, A., Hristovski, D.: Disentangling the evolution of MEDLINE bibliographic database: a complex network perspective. J. Biomed. Inform. **89**, 101–113 (2019)
25. Konopka, B.M.: Biomedical ontologies–a review. Biocybern. Biomed. Eng. **35**(2), 75–86 (2015)
26. Lipscomb, C.E.: Medical subject headings. Bull. Med. Libr. Assoc. **88**(3), 265 (2000)
27. Lowe, H.J., Barnett, G.O.: Understanding and using the Medical Subject Headings (MeSH) vocabulary to perform literature searches. JAMA **271**(14), 1103–1108 (1994). https://doi.org/10.1001/jama.1994.03510380059038
28. Lu, K., Yang, G., Wang, X.: Topics emerged in the biomedical field and their characteristics. Technol. Forecast. Soc. Chang. **174**, 121218 (2022). https://doi.org/10.1016/j.techfore.2021.121218
29. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947). https://doi.org/10.1214/aoms/1177730491
30. Mizzaro, S.: Relevance: the whole history. J. Am. Soc. Inf. Sci. **48**(9), 810–832 (1997)
31. Naderi, N., Knafou, J., Copara, J., Ruch, P., Teodoro, D.: Ensemble of deep masked language models for effective named entity recognition in health and life science corpora. Front. Res. Metrics Anal. **6**, 689803 (2021)
32. Nelson, S.J., Schulman, J.: A multilingual vocabulary project-managing the maintenance environment. MeSH Section, National Library of Medicine, Bethesda, Maryland (2007)
33. Newman, M.: Networks. Oxford university press (2018)
34. NIH: MeSH browser. https://meshb.nlm.nih.gov/treeview, Accessed 14 July 2025
35. NIH: Record types. https://www.nlm.nih.gov/mesh/intro_record_types.html, Accessed 14 July 2025
36. Park, M., Leahey, E., Funk, R.J.: Papers and patents are becoming less disruptive over time. Nature **613**(7942), 138–144 (2023). https://doi.org/10.1038/s41586-022-05543-x
37. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. Cogn. Psychol. **8**(3), 382–439 (1976). https://doi.org/10.1016/0010-0285(76)90013-x
38. Saracevic, T.: Relevance: a review of and a framework for the thinking on the notion in information science. J. Am. Soc. Inf. Sci. **26**(6), 321–343 (1975)
39. Schamber, L., Eisenberg, M.B., Nilan, M.S.: A re-examination of relevance: toward a dynamic, situational definition. Inf. Process. Manage. **26**(6), 755–776 (1990)

40. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
41. Teodoro, D., et al.: Information retrieval in an infodemic: the case of COVID-19 publications. J. Med. Internet Res. **23**(9), e30161 (2021)
42. Wang, X., et al.: Quantifying knowledge from the perspective of information structurization. PLoS ONE **18**(1), e0279314 (2023)
43. Wu, L., Wang, D., Evans, J.A.: Large teams develop and small teams disrupt science and technology. Nature **566**(7744), 378–382 (2019)

Fine-Tuning of Small Language Models for Ecological Monitoring: A Comparative Analysis Based on Perplexity

Jhon Edilberto Monroy Barrios(✉), Enrique Miguel Paco Cusi, Miguel Angel Portella Huincho, Jeymi Melanie Valdivia Eguiluz, Basurco Zapata Gerald Francis, and Fiorella Jara Rosales

Universidad de San Martín de Porres, Arequipa 04001, Peru
{jmonroyb,epacoc,miguel_portella,jvaldiviae,gbasurcoz,fjarar}@usmp.pe

Abstract. This study presents a comprehensive evaluation of fine-tuning applied to Small Language Models (SLMs) for specialization in highly specific scientific domains, with a case study in ecological monitoring. A comparative analysis was conducted between the `Llama 3.2-3B` model, adapted using a proprietary corpus on native flora, and the baseline `DeepSeek-R1-Distill-Qwen-1.5B` model. Quality was measured using the perplexity metric, yielding substantial improvements: the fine-tuned model achieved values of 1.25 and 1.60 on the `DPAV1.0` and `DPAV2.0` datasets, compared to 49.48 and 108.95 for the baseline model, demonstrating orders-of-magnitude superior contextual understanding and terminological precision.

The experimental architecture integrates environmental sensors, cloud infrastructure, a REST API, and a conversational agent based on the specialized SLM, capable of processing real-time data and generating actionable recommendations. Computational performance analysis showed that, with a moderate consumption of 3 GB VRAM and competitive inference times, the fine-tuned model is viable for deployment on mid-range hardware, enabling adoption in resource-constrained environments.

The results confirm that SLM fine-tuning is not only technically feasible but also a highly efficient strategy for creating expert AI agents in niche domains, offering a practical and scalable alternative to massive LLMs. This work sets a precedent for the development of specialized, accurate, and low-cost AI systems with direct applications in environmental conservation.

Keywords: Small Language Models · Fine-Tuning · Ecological Monitoring · Perplexity · Natural Language Processing

1 Introduction

Ecological monitoring is essential for biodiversity conservation and the sustainable management of ecosystems. However, the continuous tracking of species and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 300–313, 2026.
https://doi.org/10.1007/978-3-032-20322-9_21

habitats involves processing large volumes of data from sensors, camera traps, and passive acoustic monitoring [8]. In this context, Artificial Intelligence (AI) has emerged as a transformative tool, and in particular, language models have opened new opportunities to interpret textual data and generate contextualized recommendations with high potential for ecological science [3].

The rapid development of Large Language Models (LLMs) has demonstrated outstanding capabilities in various tasks. Nevertheless, their high computational cost, substantial memory requirements, and reliance on centralized infrastructures limit their use in field settings or on resource-constrained devices conditions that are common in environmental monitoring [11]. Given these limitations, Small Language Models (SLMs) have positioned themselves as an efficient alternative: they reduce costs and latency, facilitate local deployment, and enhance privacy and accessibility [14].

Despite these advantages, general-purpose SLMs often lack the specialized knowledge required for niche domains, such as the botany of native flora. This gap can lead to errors or "hallucinations" that compromise reliability in critical scientific applications. An effective strategy to address this challenge is fine-tuning, a transfer learning technique that adapts a pre-trained model to a specific domain using a limited amount of data. This approach, consolidated by works such as ULMFiT [5], is now a cornerstone of modern Natural Language Processing (NLP) [12].

In this article, we evaluate the effectiveness of fine-tuning applied to an SLM in the highly specialized domain of native plant monitoring [15]. To this end, we compare a `Llama 3.2-3B` model fine-tuned with a proprietary corpus (DPAV1.0) against a general-purpose `DeepSeek-R1-Distill-Qwen-1.5B` model, using perplexity as the quality metric. The results show a drastic reduction in perplexity from 49.48 to 1.25 demonstrating the ability of fine-tuning to significantly enhance the contextual understanding and accuracy of SLMs in real world ecological applications.

1.1 Previous Work

The use of language models in niche domains has gained relevance over the past decade due to their ability to process specialized terminology and generate high-precision contextual inferences. In the ecological field, [8] demonstrated how artificial intelligence can enhance environmental monitoring through the integration of data generated by sensors, camera traps, and citizen science platforms, enabling large-scale ecosystem analysis. Complementarily, the systematic review by [14] highlights that Small Language Models (SLMs) represent a viable alternative to LLMs for resource-limited environments, thanks to their lower latency, reduced energy consumption, and ease of local deployment.

In Natural Language Processing (NLP), fine-tuning has been established as a central technique for adapting pre-trained models to new domains with limited data. The foundational work of [5] introduced progressive fine-tuning as a strategy to maximize performance in text classification tasks, while demonstrated

its effectiveness in detecting diseases in tea leaves, outperforming both training from scratch and transfer learning without fine-tuning.

In the field of computer vision, [1] applied fine-tuning to pre-trained convolutional networks for hyperspectral image classification, achieving substantial accuracy improvements over conventional approaches. Similarly, [2] proposed blockwise fine-tuning of network layers, which mitigated overfitting and improved generalization in medical classification with scarce data.

Although these advances originate in different domains, they provide transferable principles such as selective layer optimization and the use of domain-specific data augmentation. These principles inspire the design of our approach to optimize SLMs for ecological monitoring, combining semantic specialization with computational efficiency.

2 Experimental Methodology

To rigorously evaluate the effectiveness of fine-tuning in small language models for the ecological domain, we designed a methodological framework that encompasses every stage from system architecture to the quantitative evaluation of the models. This section details the components of our system, the selected models, the data curation and fine-tuning process, and the metrics used for comparison.

2.1 Ecological Monitoring System Architecture

The system was designed as a platform for the realtime collection, processing, and analysis of environmental data, adopting a Cyber-Physical System (CPS) approach. The architecture, as illustrated in Fig. 1, integrates both hardware and software components to create an interactive intelligent agent, ensuring scalability and modularity in field deployment [7].

The key components are:

- **Environmental Sensors (Physical Layer):** A set of low-power sensors (temperature, humidity, light level, soil moisture) connected to an ESP32 microcontroller collects data from the plant's environment, constituting the physical-world interface.
- **Data Layer (Google Cloud):** Sensor data is transmitted via MQTT and stored in the cloud (Firebase/GCP), creating a secure and highly available historical record of environmental conditions.
- **RESTful API (FastAPI):** Acts as the system's core logic and middleware layer, managing user requests, retrieving contextual data from the cloud, and orchestrating communication with the AI model.
- **Language Model (SLM - Intelligence Core):** The intelligence core, responsible for generating analyses and recommendations based on the user's query and the sensor data, effectively translating data into actionable knowledge.

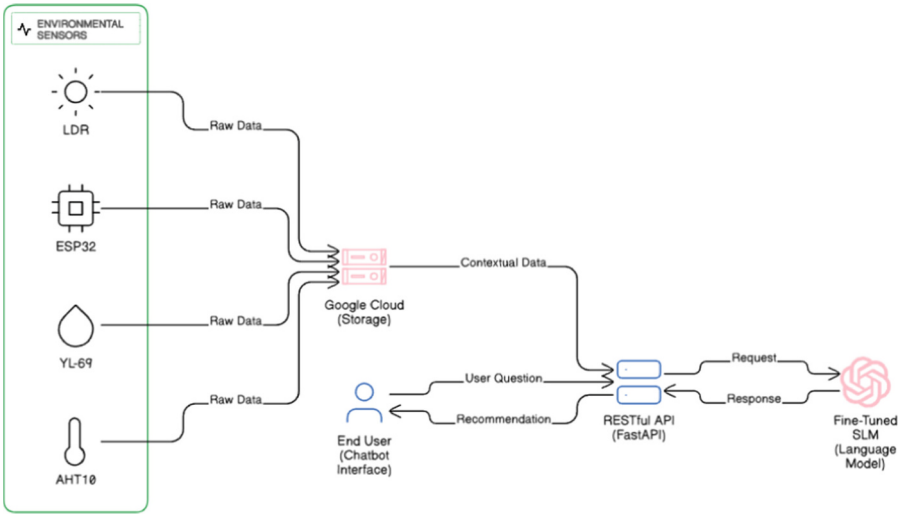


Fig. 1. System architecture diagram showing the data flow from environmental sensors to the final user interaction through the fine-tuned language model.

2.2 Model Selection and Description

For our comparative analysis, we selected three Small Language Models (SLMs) to rigorously assess the impact of domain specialization. The selection includes the target model, its pre-trained counterpart for methodological control, and an independent efficient baseline, as detailed in Table 1.

Table 1. Small Language Models (SLMs) selected for the comparative study and their roles

Model	Parameters	Role in the study
Llama 3.2-3B (Fine-tuned)	3B	Target model for specialization via Supervised Fine-Tuning (SFT)
Llama 3.2-3B (Base)	3B	**Methodological Control.** Used to isolate the performance gain attributable exclusively to domain fine-tuning (SFT)
DeepSeek-R1-Distill-Qwen-1.5B	1.5B	External Baseline. Selected to assess the specialized model against a highly efficient, parameter-optimized generalist alternative

The core of the study is the specialized model, **Llama 3.2-3B (Fine-tuned)**. The choice of its pre-trained base, **Llama 3.2-3B (Base)**, is fundamental: this initialization strategy leverages transfer learning to adapt robust,

general linguistic knowledge to the highly specific target domain, proving substantially more efficient and reliable than training from scratch.

To address the methodological concern regarding the validity of the comparison, the Llama 3.2-3B (Base) model is included as the primary control. By comparing the performance metrics (specifically Perplexity) of the Llama Base versus the Llama Fine-tuned model, we can isolate the effect of the fine-tuning process (our primary variable) from the inherent architectural capabilities.

Conversely, **DeepSeek-R1-Distill-Qwen-1.5B** serves as an external baseline. This model represents a class of highly efficient, distilled SLMs, allowing us to robustly assess whether our domain-specialized model maintains a competitive edge over smaller, optimized general-purpose models, which is crucial for evaluating its practical significance in resource-constrained environments.

2.3 Data Curation and Fine-Tuning Process

The success of fine-tuning critically depends on the quality and specificity of the adaptation dataset. We created two proprietary and custom datasets in JSON format, structured as question–answer pairs, exclusively focused on native Peruvian flora. The development of this high-fidelity, domain-specific corpus is essential for overcoming the knowledge scarcity problem observed in generalist models when tackling highly specialized scientific fields [9].

- **DPAV1.0 (Dataset Plantas Autóctonas):** Contains 1,500 question-answer pairs ($\approx$ 250k tokens). This dataset focuses on general knowledge about native plants, including descriptions, care guidelines, and species characteristics, forming the foundational basis of specialized botanical terminology.
- **DPAV2.0 (Dataset Plantas Autóctonas Ambiente):** Extends DPAV1.0 with 500 pairs ($\approx$ 90k tokens). This set is critical as it integrates questions that require the analysis of structured, real-time environmental sensor data (e.g., "What is the status of my plant with a humidity level of 57.1%?"). The format for these examples was strictly designed to teach the model **data grounding**, forcing it to integrate numerical input (JSON snippet from the RESTful API) directly into a natural language analysis. This step is crucial for transforming the model from a linguistic predictor into a **contextual data interpreter**.

The supervised fine-tuning (**SFT**) process was applied exclusively to the pre-trained **Llama 3.2-3B model**, following the procedure described in Algorithm 1. We strategically employed **Full Fine-Tuning** rather than Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA. For a Small Language Model (SLM) of only 3 billion parameters, the computational cost of full fine-tuning is acceptable, and this approach allows us to **maximize domain knowledge embedding** by updating all model weights, which is vital for achieving high fidelity in a specialized scientific domain. While PEFT methods are essential for managing the prohibitive memory and time demands of massive Large

Language Models (LLMs) (>50B parameters), we prioritized deep and comprehensive parameter adaptation over marginal resource savings in our constrained setting. The entire dataset was processed over 3 epochs with a batch size of 4, a configuration optimized to prevent catastrophic forgetting while ensuring deep adaptation to the new domain's syntax and semantics.

Algorithm 1 Supervised Fine-Tuning (SFT) for Domain Adaptation

Require: Pre-trained Language Model $\mathcal{M}_{\text{base}}$, Fine-tuning Dataset $\mathcal{D}_{\text{SFT}}$, Learning Rate η, Number of Epochs E, Batch Size B
Ensure: Fine-tuned Language Model $\mathcal{M}_{\text{fine-tuned}}$
1: $\mathcal{M}_{\text{fine-tuned}} \leftarrow \mathcal{M}_{\text{base}}$ ▷ Initialize fine-tuned model with base weights
2: **for** $e = 1$ to E **do**
3: Shuffle $\mathcal{D}_{\text{SFT}}$
4: Split $\mathcal{D}_{\text{SFT}}$ into mini-batches $\mathcal{B}$
5: **for** each batch $B \in \mathcal{B}$ **do**
6: Compute predictions $\hat{Y}$ on B using $\mathcal{M}_{\text{fine-tuned}}$ ▷ Forward Pass
7: Compute Loss $\mathcal{L}(Y, \hat{Y})$ using Cross-Entropy
8: Compute gradient $\nabla_{\Theta}\mathcal{L}$ ▷ Backward Pass
9: Update model parameters Θ via AdamW: $\Theta \leftarrow \Theta - \eta \cdot \nabla_{\Theta}\mathcal{L}$ ▷ Optimization step
10: **end for**
11: **end for**
12: **Return** $\mathcal{M}_{\text{fine-tuned}}$

The specialization process for our **Small Language Model (SLM)** is governed by Algorithm 1. This core procedure is fundamentally an iterative, gradient-based error minimization process. The algorithm takes as input the pre-trained base model ($\mathcal{M}_{\text{base}}$) and our curated specialized corpus ($\mathcal{D}_{\text{SFT}}$), alongside key hyperparameters: the **AdamW optimizer** with a fixed learning rate of $\eta = 2e-5$ (selected for stable convergence), the number of epochs ($E = 3$), and a batch size of $B = 4$. In each epoch, the dataset is shuffled (L3) to mitigate sequential ordering bias. For every mini-batch, a **Forward Pass** (L6) obtains the predicted token sequence, and the **Cross-Entropy Loss** function ($\mathcal{L}$) quantifies the discrepancy. Subsequently, the Backward Pass (L8) computes the gradients, enabling the AdamW optimizer to adjust the full set of model parameters (Θ) (L9). This continuous adjustment cycle efficiently transfers the complex domain knowledge into the model's weights, ultimately transforming the general-purpose SLM into a specialized knowledge retriever and data interpreter.

2.4 Evaluation Metrics

To rigorously quantify improvements in model quality and efficiency, we used two categories of metrics.

Language Quality Metric: Perplexity. Perplexity (PPL) is the primary intrinsic metric for evaluating the quality of a language model. It measures how well a probabilistic model predicts a sample. This metric was strategically selected because a lower PPL score directly quantifies the model's success in modeling the specific terminologies and semantic structures inherent to the niche ecological domain, thus serving as a proxy for contextual accuracy and reduction of'hallucination.' Lower perplexity indicates that the model is less "perplexed" by the test text and, therefore, models it more effectively. It is calculated as the exponential of the average cross-entropy. For a sequence of tokens $W = (w_1, w_2, \ldots, w_N)$, perplexity is defined as:

$$PPL(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i|w_1, \ldots, w_{i-1}; \theta)\right) \tag{1}$$

where N is the total number of tokens in the sequence and $p(w_i|w_1, \ldots, w_{i-1}; \theta)$ is the probability of token w_i predicted by the model with parameters θ.

Computational Performance Metrics. To assess the practical feasibility of the models for deployment in resource-constrained environments, we measured the following performance indicators:

- **Inference Time (seconds):** The time required to generate a complete response, measured on both CPU and GPU.
- **VRAM Usage (GB):** The amount of video memory required to load and run the model on a GPU.

2.5 Experimental Setup

All experiments were conducted in a controlled hardware environment to ensure the reproducibility and comparability of results. The system specifications are summarized in Table 2.

Table 2. Hardware specifications used for performance and evaluation experiments.

Component	Specification
GPU	NVIDIA GeForce RTX 2060
Minimum RAM	16 GB
Virtual CPUs (vCPUs)	4
Operating System	Ubuntu 18.04 or later

3 Results and Analysis

This section presents the empirical results of our experiments, comparing the fine-tuned Llama 3.2-3B model with the baseline DeepSeek-R1-Distill-Qwen-1.5B model. The analysis focuses on two main aspects: model quality, evaluated through the perplexity metric, and computational efficiency, assessed in terms of performance and resource consumption.

3.1 Model Quality Evaluation: Perplexity Analysis

Perplexity (PPL) is the core metric used to assess a language model's ability to coherently predict text within a specific domain. Lower PPL scores indicate greater model confidence and superior accuracy in modeling the domain-specific language.

Table 3 reports the perplexity results for the three models, evaluated on our custom datasets, DPAV1.0 and DPAV2.0. The inclusion of the Llama 3.2-3B (Base) model serves as the methodological control to isolate the performance gain resulting from the fine-tuning process.

Table 3. Comparison of Perplexity (PPL) scores across the specialized model, the methodological control, and the external baseline. Lower PPL indicates superior domain language modeling capability.

Model	Dataset	Perplexity
Llama 3.2-3B (Fine-tuned)	DPAV1.0 (General Knowledge)	**1.25**
	DPAV2.0 (Environmental Data)	**1.60**
Llama 3.2-3B (Base)	DPAV1.0 (General Knowledge)	62.50
	DPAV2.0 (Environmental Data)	135.20
DeepSeek-R1-Distill-Qwen-1.5B	DPAV1.0 (General Knowledge)	49.48
	DPAV2.0 (Environmental Data)	108.95

The most significant finding of our controlled study is the drastic, orders-of-magnitude reduction in perplexity achieved after the supervised fine-tuning (SFT) process. The PPL score dropped from 62.50 (Llama Base) to 1.25 (Llama Fine-tuned) on DPAV1.0, representing a reduction factor of approximately 50. Similarly, on the more complex DPAV2.0, PPL fell from 135.20 to 1.60.

This result directly resolves the methodological concern. Since both the base and fine-tuned models share the exact same architecture, the observed performance leap is attributable solely to the SFT process and the specialization achieved with the proprietary domain corpus.

The specialized model's scores (1.25 and 1.60) not only validate the effectiveness of fine-tuning but also demonstrate a domain language modeling capability significantly superior to both the methodological control and the external

Table 4. Comparative performance of the models on GPU (NVIDIA GeForce RTX 2060) and CPU. Inference time is reported in seconds, and VRAM consumption in gigabytes.

Model	GPU (NVIDIA GeForce RTX 2060)		CPU (4 vCPUs)	
	Inference Time (s)	VRAM (GB)	Inference Time (s)	VRAM (GB)
Llama 3.2-3B (Fine-tuned)				
Domain-Specific Questions	6	3.0	45	N/A
General Questions	5	3.0	25	N/A
Llama 3.2-3B (Base)				
Domain-Specific Questions	6	3.0	45	N/A
General Questions	5	3.0	25	N/A
DeepSeek-R1-Distill-Qwen-1.5B				
Domain-Specific Questions	7	2.7	42	N/A
General Questions	5	2.7	23	N/A

DeepSeek baseline. This transformation establishes the specialized Llama 3.2-3B as a domain expert capable of accurately understanding and predicting the terminology and semantic structures of ecological monitoring.

3.2 Computational Performance Analysis

Beyond quality metrics, we assessed the practical feasibility of the models by measuring their computational performance on consumer-grade hardware, which is a key requirement for deployment in resource-constrained environments. Table 4 compares inference times and Video Random Access Memory (VRAM) usage across the specialized model, the external baseline, and the methodological control (Llama Base).

The performance results reveal important trade-offs between specialization and latency. As expected, GPU inference is significantly faster for all models, with execution times ranging from 5 to 7 s compared to 23 to 45 s on the CPU. Importantly, the Llama 3.2-3B Fine-tuned model despite its specialization and larger size than DeepSeek—maintains GPU inference times (5–6 s) and VRAM consumption (3.0 GB) virtually identical to the Llama 3.2-3B Base control model. This confirms that the Supervised Fine Tuning (SFT) process does not introduce significant computational overhead to the base architecture. Compared to the smaller DeepSeek baseline, the specialized Llama model requires only marginally more resources (0.3 GB VRAM) while achieving an orders-of-magnitude superiority in domain quality (as shown by PPL). These findings confirm the practical viability of the fine tuned model for deployment on mid-range hardware and in resource-constrained environments.

3.3 Structured Qualitative Analysis

To move beyond anecdotal evidence and provide robust support for the quantitative perplexity findings, we conducted a structured qualitative evaluation [4]. This analysis focuses on two critical factors: contextual coherence and resistance to text degeneration, particularly when integrating sensor data (DPAV2.0). To ensure comparability, all models were tested with the exact same prompts.

Table 5 presents a side-by-side comparison of verbatim responses from the fine-tuned model, the external baseline, and the base control model, accompanied by an expert assessment.

Table 5. Structured, side-by-side qualitative comparison of model responses based on identical prompts.

Prompt Type	Llama 3.2-3B (Fine-tuned) Response (Verbatim)	Llama 3.2-3B (Base) Response (Verbatim)	DeepSeek-R1 Response (Verbatim)	Expert Assessment
DPAV2.0 (Sensor Data Integration)	"Data for 'Plant 1' on 2025-01-28: Humidity: 57.1%... Analysis: The plant is in a healthy state and does not require immediate intervention"	"I cannot access real-time or historical sensor data. Please provide the current humidity level"	"The status is fine. Do you need more information about the day today? Yesterday, I feel great..."	High Factual Precision. Correctly integrates and interprets internal structured data, which is essential for ecological monitoring
General Domain (Contextual Coherence)	"The importance lies in preserving the local biodiversity, specifically the native Peruvian flora of the Arequipa region, vital for the ecosystem's resilience"	"Environmental awareness is important everywhere. It helps people recycle, reduce pollution, and protect large animals"	"Why is environmental awareness important in Arequipa? How can we improve the quality of life in Arequipa? What can they do if it destroys us in the water?????????????..."	Superior Domain Specificity and Coherence. The Fine-tuned model maintains terminological focus and avoids the text degeneration observed in the generalist models

This structured analysis directly supports the quantitative PPL findings. The Llama Fine-tuned model consistently produces useful, coherent, and reliable responses, whereas both the Llama Base (control) and the DeepSeek baseline fail: the Llama Base lacks the necessary domain knowledge to interpret specific requests, and DeepSeek exhibits high levels of text degeneration and low contextual relevance. The transformation from the Llama Base's inability to interpret data to the Fine-tuned model's high factual precision underscores the effectiveness of our specialization strategy.

4 Discussion

The results presented in the previous section provide strong empirical support for our central hypothesis: domain-specific *fine-tuning* is a highly effective strategy for developing expert AI agents, especially when compared against general-purpose models, including those optimized for efficiency like distilled models. The drastic reduction in perplexity from ∼60 (Llama Base) to an almost perfect 1.25 on the DPAV1.0 dataset represents a qualitative transformation in the model's domain competence, rather than a mere incremental improvement.

4.1 Implications of Perplexity Results

The core finding of our study is the disproportionate impact of domain specialization on model quality, rigorously demonstrated by the Perplexity (PPL) analysis. The stark contrast between the Llama 3.2-3B (Base) control model and the Fine-tuned Llama 3.2-3B model is the primary driver of this conclusion.

The PPL scores dropped from ∼60 (Base) to 1.25 (Fine-tuned) on DPAV1.0. Since both models share an identical architecture, this orders-of-magnitude improvement is attributable solely to the Supervised Fine-Tuning (SFT) process. This directly resolves the methodological concern raised in the review process,

empirically supporting the specialization strategy as a robust and highly reliable mechanism for competence acquisition, independent of the base model's inherent advantage.

This finding implies that fine-tuning imparts more than just specialized vocabulary; the model successfully learns the semantic relationships, syntactic structures, and inherent data patterns of the ecological domain. This is functionally evidenced by the specialized model's ability to correctly integrate complex sensor data into coherent, contextually accurate responses—a task the generalist models (Llama Base and DeepSeek) fundamentally failed to accomplish, often resulting in text degeneration or severe incoherence.

Our work contributes a concrete and methodologically controlled case study in the scientific domain of ecological monitoring. It strongly supports the hypothesis that a relatively modest Small Language Model (3B parameters) can be reliably transformed into a reliable domain expert—a capability that is essential for deploying accessible, high-fidelity AI systems in resource-constrained and specialized environments, such as those prevalent in Latin America.

4.2 The Paradigm of Efficient SLMs Vs. Massive LLMs

Our computational performance results reinforce the growing argument in favor of Small Language Models (SLMs) for real-world applications. The fine-tuned Llama 3.2-3B, while larger than its 1.5B counterpart, operates with a VRAM footprint of only 3 GB and delivers competitive inference times on GPU. This positions it as a viable solution for deployment on mid-range hardware, democratizing access to specialized AI without reliance on costly massive-model APIs.

This study illustrates a principle of "expert efficiency": it is more practical and effective to deploy a small, fine-tuned "expert" for a specific task than a massive "generalist" that may lack the required niche knowledge. This is a critical consideration for fields such as environmental sciences, where solutions must be field-deployable and resource-efficient.

4.3 Study Limitations

As with any research, our study presents certain limitations that open avenues for future work.

- **Model Scope:** The comparison focused on two specific architectures. Extending the analysis to include other prominent SLMs, such as Microsoft's Phi family [10] or Google's Gemma [13], as well as exploring parameter-efficient fine-tuning (PEFT) techniques such as LoRA, could further reduce the computational requirements of adaptation [6].
- **Human Expert Validation (CRITICAL):** While perplexity is a robust intrinsic metric that measures domain fluency, it does not fully capture the usefulness or factual accuracy of the generated recommendations. A formal evaluation by domain experts in ecology is a crucial and necessary next step to fully validate the system's practical reliability and quality of advice.

- **Dataset Size:** Our DPAV1.0 and DPAV2.0 datasets, while effective, remain relatively modest in size. Investigating the impact of scaling these datasets on model performance would be a fruitful area for future research.
- **Hardware Constraints:** Although the model is resource-efficient, tests indicate a minimum requirement of 16 GB of RAM for CPU-based operation, which limits deployment on extremely resource-constrained edge devices.

5 Conclusion and Future Work

This study strongly supports the effectiveness of **Supervised Fine-Tuning (SFT)** for specializing a Small Language Model (SLM) in the complex, data-grounded domain of ecological monitoring. Through a rigorous, methodologically controlled quantitative comparison (Llama Base vs. Fine-tuned), we showed that a specialized Llama 3.2-3B model achieved a substantial, orders-of-magnitude reduction in perplexity. This metric directly translates into qualitatively superior, contextually relevant, and factually coherent responses compared to generalist baselines, particularly when integrating structured sensor data.

Our findings empirically validate that SLM specialization is a powerful and efficient strategy for creating domain-expert AI agents. Crucially, this work demonstrates a practical, low-cost alternative to massive proprietary models for niche scientific applications. By successfully adapting a 3B-parameter model, we contribute to the growing literature supporting the use of smaller, more deployable architectures, addressing the high computational costs and privacy concerns associated with large language models (LLMs). This efficiency is particularly relevant for **emerging regions** or institutions operating under resource constraints, where on-premise deployment of specialized AI is critical for environmental conservation efforts. The system's ability to interpret real-time ecological data and convert it into actionable knowledge offers a tangible foundation for future automated decision-support tools in the field.

Building on these foundational results, we propose the following critical directions for future research:

- **Domain Expert Validation (Critical Priority):** The next immediate step requires implementing a robust human evaluation protocol involving certified ecologists and botanists. This will rigorously assess the **factual accuracy**, safety, and practical utility of the system's recommendations, providing the necessary external validation to transition the model from an academic proof-of-concept to a reliable field-deployable asset.
- **Extensive SLM Benchmarking and Efficiency Trade-offs:** Conduct a broader comparative study including diverse SLM architectures (e.g., Phi-3, Gemma-2B) and a structured analysis of the trade-offs between **Full Fine-Tuning** and various Parameter-Efficient Fine-Tuning (PEFT) methods. The goal is to conclusively identify the optimal, most resource-efficient combination of model and technique for maintaining high fidelity in scientific domains.

- **Longitudinal Field Deployment and Edge Computing:** Deploy the system in a real ecological environment for long-term monitoring. This will test its robustness, reliability, and tangible impact on conservation practices, potentially exploring deployment optimization using **edge computing** techniques to minimize latency and cloud reliance in remote areas.
- **Hybridization with Causal and Predictive Models:** Integrate the language model with traditional numerical predictive models (e.g., time-series forecasting). This will enable the SLM not only to diagnose the current state of the plant but also to perform **causal reasoning** and forecast its future evolution under varying environmental scenarios, moving the agent from reactive diagnosis to proactive management.

Ultimately, this work lays the groundwork for developing more accessible, efficient, and specialized AI tools capable of supporting scientists and practitioners in the critical task of environmental conservation and accelerating the adoption of AI expertise in resource-constrained scientific disciplines.

Acknowledgments. This work was supported by the *Programa Nacional de Investigación Científica y Estudios Avanzados* (ProCiencia), an executing unit of CONCYTEC. The research was conducted within the framework of the project *"GREEN CLOUD: Sustainable Plant Ecosystem through the Use of Sensors Managed by a Programmable Electronic Board to Raise Awareness for the Care of Native Plants in the Arequipa Region"*, under Agreement No. PE501086972-2024.

References

1. Baird, H., et al.: Fine-tuning pre-trained cnns for hyperspectral image classification. Remote Sens. **14**(9), 2011 (2022)
2. Barakat, A., Huang, X.: Block-wise fine-tuning of convolutional neural networks for medical image classification. Comput. Biol. Med. **152**, 106391 (2023)
3. Eco, A., Scientist, A.: The potential of language models in ecological science. In: Proceedings of the Conference on AI for Environmental Sciences (2022)
4. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018)
5. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 328–339. Association for Computational Linguistics (2018)
6. Hu, J., et al.: Lora: low-rank adaptation of large language models. In: ICLR, vol. 1, no. 2, p. 3 (2022)
7. Lee, E.A.: Cyber-physical systems: design challenges. IEEE Proc. Cyber-Phys. Syst. **3**(1), 5–13 (2015)
8. McClure, E.C., Sievers, M., Brown, C.J., et al.: Artificial intelligence meets citizen science to supercharge ecological monitoring. Patterns **1**(7), 100109 (2020)
9. Ramdan, A., Heryana, A., Arisal, A., Budiarianto, R., Kusumo, S., Pardede, H.F.: Transfer learning and fine-tuning for deep learning-based tea diseases detection on small datasets. In: 2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), pp. 206–211 (2020)

10. Mercy Ranjit, Shaury Srivastav, and Tanuja Ganu. Radphi-3: Small language models for radiology. *arXiv preprint* arXiv:2411.13604, 2024
11. Smith, J.: Computational costs and limitations of large language models. J. Artif. Intell. Res. (2023)
12. Author Survey. A survey on fine-tuning techniques in modern nlp. Computational Linguistics (2021)
13. Team, G., et al.: Gemma: open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
14. Valipour, H., Fath-Allah, A., Motie, S., et al.: A systematic literature review on small language models. Information **15**(7), 389 (2024)
15. Yang, Yu., Mishra, S., Chiang, J., Mirzasoleiman, B.: Smalltolarge (s2l): scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. Adv. Neural. Inf. Process. Syst. **37**, 83465–83496 (2024)

A Variance-Invariance Approach to Better Word Embeddings

Zainab Albujasim[1(✉)] and Diana Inkpen[2]

[1] Dakota State University, 820 Washington Ave N, Madison, SD, USA
zainab.albujasim@dsu.edu
[2] University of Ottawa, 800 King Edward Avenue, Ottawa, ON, Canada
diana.inkpen@uottawa.ca

Abstract. Word embeddings continue to be used in natural language processing for their efficiency, interpretability, and ease of integration into a wide range of applications. While effective at capturing distributional semantics from large unstructured corpora, they often miss fine-grained lexical relationships such as synonymy, antonymy, and hierarchy. This paper introduces a lightweight post-processing framework designed to enrich pre-trained word embeddings by incorporating structured lexical knowledge from resources like WordNet. Our approach leverages a self-supervised training objective based on variance and invariance principles to adjust word vector relationships according to their semantic roles. This enriched structure is then propagated back to the original embedding space without requiring retraining. Experimental results show that our method consistently improves word similarity performance by an average of 8%, demonstrating the value of integrating external lexical knowledge into pre-trained embeddings for enhanced semantic representation.

Keywords: Post-processing · word embedding · variance · invariance

1 Introduction

Language processing has become one of the most important pillars of modern technology. Word embeddings are a cornerstone technique in natural language processing (NLP), enabling the representation of words as numerical vectors that capture semantic and syntactic relationships. Although word embeddings provide strong performance for many NLP tasks, they mainly encode semantic and syntactic information. For instance, they can effectively identify a semantic association between "cat" and "tiger," but determining the specific nature of the relationship—such as both being felines—can be more challenging. Meanwhile, dictionaries and lexical resources such as WordNet and BabelNet are abundant and ripe with features that can be used to boost the quality of word embeddings. In recent years, there has been an emphasis on fine-tuning word embedding models by integrating semantic knowledge such as synonyms and antonyms. Post-processing injects external knowledge without the need for expensive training,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 314–326, 2026.
https://doi.org/10.1007/978-3-032-20322-9_22

and an effective post-processing technique will create word embeddings better suited to particular NLP tasks.

Post-processing refers to a set of techniques applied to word embedding models to enhance their quality. These techniques aim to improve the similarity between related words or incorporate external knowledge from other resources. Standard post-processing methods include dimensionality reduction, normalization [1–3], fine-tuning [4,5], and clustering [6,7]. Additionally, several studies have demonstrated that autoencoder models can enhance the quality of pretrained embeddings without discarding important information. For instance [8,9], show autoencoders can increase the similarity between related words while retaining all the embedding components.

The standard post-processing algorithms usually employ two types of techniques to maintain the information carried by the original vectors while incorporating new information from lexical resources. Attract-Repel [10] and retrofitting models [11] control the distance between the original and the post-processed vector while the models from [12,13] aim to preserve the vector neighborhoods.

An extensive body of research addresses the augmentation of word embeddings with extra morphological, syntactic, semantic, and domain knowledge. The resulting techniques have generally been used to improve word embedding quality, address domain knowledge to facilitate domain-specific NLP, or handle data sparsity when the target text is small, or the target words are rare. The knowledge added to word embedding might be textual knowledge, morphological knowledge, or categorical knowledge.

Textual knowledge is the type of knowledge that might be found on Wikipedia or in general or domain-specific dictionaries. Textual descriptions often contain words that are closely related to a target word, and the objective functions of word embedding models can be used to integrate such knowledge since the embedding of a target word should be close to the embedding of its related words.

Morphological knowledge encapsulates information related to the basic components of words such as roots, prefixes, and suffixes. Morphological knowledge can be leveraged to identify semantically related words such as words with the same root, and this can allow for the capture and generalization of morphologically complex and out-of-vocabulary (OOV) words [14].

Categorical knowledge incorporates the semantic and syntactic properties of words. Words can be classified according to their meaning or semantics into categories such as concept, semantic type, or semantic role. These words, known as synonyms, are grouped into sets called synsets and share a common meaning. Syntactic information includes properties like parts of speech, and to integrate such information, an objective function is used to minimize the distance between word vectors belonging to the same category. This increases the likelihood of predicting target categories [14].

Relation knowledge [4,15] encodes the relationship between words, which is often expressed as a triplet (head H, relation R, tail T). This knowledge is injected into an embedding by adding a constraint requiring that a representation of T should be close to the sum of a representation of H and a representation of R.

In our work, we propose a novel approach that uses post-processing techniques that incorporate relations extracted from WordNet in order to enhance the quality of word embeddings. We propose to use self-supervised learning based on variance and invariance [16] to align vectors for pairs of words related by specific relationships. We then propagate this information back to the original embedding matrix, and experimental results show that our model consistently improves the semantic similarity of word embeddings by an average of 8%.

2 Related Work

Recent research has explored various techniques to refine word embeddings, focusing on geometric properties and how adjustments to their structure can enhance performance in downstream NLP tasks. A key insight comes from Mu et al. [17], who address the "hubness" problem—where word embeddings cluster too tightly in high-dimensional space—by removing the common mean vector. This simple step, they found, significantly improved the quality of embeddings, leading to better results on a range of tasks. Another approach to improving word embeddings involves using autoencoders, which focus on learning more efficient, compact representations without removing any components from the original embeddings. Kaneko et al. [8] show that autoencoders can refine pre-trained embeddings by optimizing them through unsupervised training, preserving essential information while enhancing performance on downstream tasks. This method improves the quality of embeddings by capturing more meaningful features through reconstruction, rather than discarding any dimensions.

Post-processing methods also vary based on the lexical resources they use. Some techniques utilize dictionaries and improve embedding by incorporating new definitions or words with similar meanings. Others use resources such as WordNet or BabelNet to extract sets of synonyms, antonyms, hypernyms, or homonyms. A good example of this approach is Lexical Entailment Attract-Repel (LEAR) [18]. Still, other methods involve paraphrasing databases [19].

The model from [4] used relational information found in semantic lexicons under the assumption that linked words should be similarly represented, while the model of [5] used a deep neural network to fine-tune word vectors by adding constraints on synonyms and antonyms, which yielded a significant improvement for topical similarity datasets.

The AutoExtend model [20] was designed to learn embeddings for synsets and lexemes from WordNet lexical without the need for additional training corpora. The model assumes that the vector of a word is the sum of its lexemes and that a synset is the sum of lexemes. The HyperVec model [21] proposed to learn embeddings that capture hypernym-hyponym relationships. The model reflects lexical knowledge by adjusting the word vectors associated with linguistic constraints [22].

The Dict2Vec model [12] proposed a method for learning word embedding using lexical dictionaries which are constructed using both corpus and dictionary entries. Dict2vec is based on the Skip-Gram model. In contrast to Skip-Gram,

which uses co-occurrence statistics between words in the corpus to learn embeddings, Dict2vec uses structured information from a lexical dictionary. CPAE [23] is a model that uses an LSTM-based autoencoding model to learn word representations from dictionary definitions. Its goal was for words with similar definitions to have similar embeddings.

3 Method

In this section, we will explain the proposed method in detail. It is motivated by the current trend involving self-supervised learning architectures [24,25] and the use of joint-embedding learning and Siamese networks [24,26]. Our model uses a symmetric neural network architecture consisting of two autoencoders, each of which is composed of an encoder followed by a decoder.

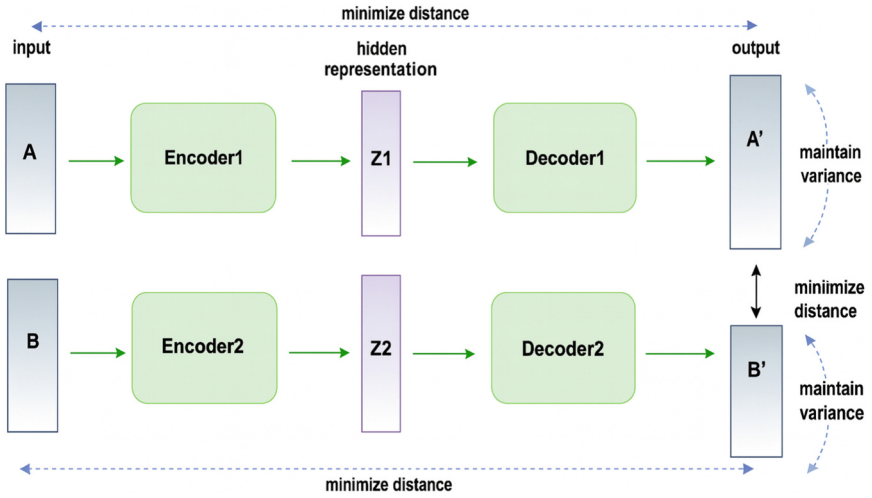


Fig. 1. Model architecture: Siamese autoencoder with variance-invariance regularization. Given a pair of input words A and B, each word is passed to the autoencoder model, and the loss function is computed at the decoder layer. The loss function consists of a variance term for the independent maintenance of the variance of each dimension and an invariance term for the minimization of the distance between the two word vectors. The autoencoder loss function also ensures that a new representation does not diverge too much from the original representation.

The decoder computes a loss function that includes variance and invariance terms in addition to the loss function of the autoencoder. The function serves to independently constrain the variance of the embeddings along each dimension and to bring the representations of pairs of words closer together through the use of a standard mean-squared Euclidean distance. Figure 1 shows the architecture of the proposed model.

After training the model, we extract the new embedding from the hidden representation of the encoder layer. We proceed to use the new, learned embedding to learn a transformation matrix and then propagate the new information to the original embedding matrix. More precisely, given a pair of word embeddings (A_i, B_j) representing words sampled from the training dataset, the vectors are input to each autoencoder in the Siamese network.

An autoencoder neural network is a neural network for which the output matches the input. Every autoencoder model consists of an encoder phase E and a decoder phase G. In our case, the encoder and decoder each consist of one layer with a $tanh()$ activation function. To train an autoencoder, we need to optimize the parameters θ to minimize the reconstruction loss between the given dataset $\mathcal{D}$ and the reconstructed dataset $\mathcal{D}'$. The objective function is thus given by

$$L(\theta) = L(\mathcal{D}, \mathcal{D}') = L(\mathcal{D}, G(E(\mathcal{D}))) \tag{1}$$

where the loss function $\mathrm{L}(\cdot, \cdot)$ is defined as the squared Euclidean distance and the loss function $L(\theta)$ is the average loss over all the word embedding samples and their reconstructions.

Variance regularization independently constrains the variance of the embeddings in each dimension. Specifically, we follow [16] to calculate a hinge loss from the standard deviations along batch dimensions. The invariance incorporates a standard mean-squared Euclidean distance and is used to learn parameter values that reduce the distance between word embedding pairs.

The word pairs are processed in batches denoted $A = [a_1, \ldots, a_n] \in R^{n \times d}$ and $B = [b_1, \ldots, b_n] \in R^{n \times d}$ with each composed of n vectors of dimension d. We denote by A_i the i^{th} vector in A and by $A_{:,j}$ the vector composed of the coordinate values for dimension j from each vector in A. Regularization of variance v is defined as a hinge loss on standard deviation as implemented in [16].

$$v(A) = \frac{1}{d} \sum_{j=1}^{d} \max\left(0, \alpha - \sqrt{\mathrm{Var}\left(A_{:,j}\right) + \epsilon}\right) \tag{2}$$

where α is the standard deviation target value set to 0.1, ϵ is a small scalar that prevents numerical instability, and $\mathrm{Var}(x)$ is the variance estimated in this case by

$$\mathrm{Var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

with n the size of x and $\bar{x}$ the mean of x. By using this criterion, each variance dimension within the current batch will take the value of α by force.

We define the invariance term s between A and B as the mean-squared Euclidean distance between each pair of word vectors [16].

$$s\left(A, B\right) = \frac{1}{n} \sum_{i} \|A_i - B_i\|_2^2 \tag{4}$$

The overall loss function is then defined as:

$$\mathcal{L} = \ell_1 + \ell_2 + \lambda s(\Lambda, B) + \mu\{v(\Lambda) + v(B)\} \tag{5}$$

where ℓ_1, ℓ_2 are the loss functions for the two autoencoders, v is the variance term, and s is the invariance term. λ and μ are hyperparameters to balance the loss function.

After training, the hidden representation from the encoder phase is used as the new representation. These embeddings are used to learn a matrix transformation that propagates information from the new embedding space back to the original embedding space. We follow [27] and use the transformation to learn a mapping function from the specialized space to the original embedding space by computing.

$$\min_W \sum_{i=1}^{n} \|Wx_i - z_i\|^2 \tag{6}$$

.

Studies such as [28,29] demonstrate that orthogonality constraints on the linear operator improve performance, preserve the dot products of vectors, and ensure the isotropy of the Euclidean space. The mapping we use is denoted by

$$W \leftarrow (1+\beta)W - \beta\left(WW^T\right)W \tag{7}$$

where β is 0.001. The network is trained using the Adam optimizer with a learning rate of 0.0001 and dropout rate 0.3 with epochs 20 and batch size 256.

4 Experiment

In this section, we evaluate our method's performance on semantic similarity using the semantic similarity benchmark dataset [30]. We employ three prominent, pre-trained word embedding models: Word2Vec with 300 dimensions trained on a Google News corpus [31]; GloVe with 300 dimensions trained on a 2014 Wikipedia corpus and the Gigaword 5 corpus, which has 400,000 vocabulary [32]; and fastText with 300 dimensions trained on one million words from Common Crawl [33].

Table 1. Average Spearman's correlation of the original embeddings and of those from our proposed model on a similarity task using only synonyms or using all extracted relations.

Model	WS -353 -ALL	MTurk -771	WS -353 -REL	YP -130	RW -STAN -FORD	WS -353 -SIM	VE -RB -143	MTurk -287	SIM -LEX -999	SIM -MC -30	RG -65	MEN -TR 3k
fastText-original	70.84	71.01	65.00	51.87	52.33	81.02	46.27	70.50	45.00	83.63	84.51	79.06
ours-only-synonyms	75.12	75.98	71.84	68.83	57.13	81.85	49.05	72.03	52.25	85.99	86.83	83.08
ours-all	**75.69**	**76.84**	**72.31**	**70.48**	**57.86**	**82.90**	**49.02**	**74.03**	**52.90**	**87.48**	**86.46**	**83.95**
GLoVe-original	60.54	65.01	57.26	56.13	41.18	66.38	30.51	63.32	37.05	70.26	76.62	73.75
ours-only-synonyms	66.35	69.19	63.13	60.37	46.72	72.08	34.33	64.20	44.49	76.80	80.72	78.80
ours-all	**66.95**	**69.43**	**63.37**	**63.36**	**46.97**	**72.81**	**36.01**	**64.8**	**43.63**	**77.85**	**80.48**	**78.77**
Word2Vec-original	69.18	67.13	62.19	55.9	53.42	77.71	49.73	68.20	44.20	78.80	74.98	77.08
ours-only-synonyms	68.59	69.36	62.19	64.19	54.84	76.44	50.65	66.86	47.10	79.14	78.58	79.05
ours-all	**70.50**	**70.21**	**64.59**	**64.34**	**54.87**	**78.10**	**52.82**	**69.471**	**47.92**	**80.76**	**78.83**	**79.66**

4.1 External Resources

We used WordNet lexical resources to build a training dataset and to extract pairs of words such as synonyms that exhibit specific relationships.

4.2 WordNet

WordNet [34] is one of the most popular NLP semantic resources. It can be represented as a graph in which nodes represent words or concepts and edges represent relationships. The basic structural unit of WordNet is a synset that represents a group of words that have similar meanings or that are more generally connected through conceptual-semantic and lexical relationships.

More formally, WordNet is a semantic network with nodes representing synsets like hypernyms or meronyms and edges representing lexical or semantic relations connecting them. It currently consists of 155,327 words and 117,979 synsets. WordNet is also a multilingual resource that extends to languages other than English through the multilingual Wordnet project [35].

For our experiment, we selected 767,306 synonym, meronym, and homonym word pairs and filter down to 186,473 by removing all pairs that are absent from the pre-trained word embedding, that consist of more than two words, or that have identical first and second words. Most of the resulting word pairs consist of synonyms.

4.3 Semantic Similarity

SentEval [36] is a centralized evaluation framework introduced with the goal of providing a standard set of evaluations based on community consensus. It generates sentence-level embeddings by computing the means of all word vectors in a sentence. SentEval has a fixed number of standard parameters to minimize discrepancies in reported results. The toolkit is implemented in Python and supports a simple interface to assist in downloading datasets. It covers a wide range of NLP tasks, including binary and multi-label classification, textual similarity, and entailment. For sentence classification, SentEval includes datasets such as Sentiment Analysis (MR) [37], Question Type (TREC) [38], Product Review (CR) [37], Subjectivity/Objectivity (SUBJ) [37], and Opinion Polarity (MPQA) [37]. It also contains datasets for entailment and semantic relatedness tasks, such as the SICK dataset for entailment (SICK-E), semantic relatedness (SICK-R), and the STS benchmark dataset which includes STS 2012-2016 datasets([39–43]), SICK-Relatedness dataset (SICK-R) [44], and the STS benchmark (STS-B) [41]. Additionally, SentEval includes a dataset for paraphrase detection, the Microsoft Research Paraphrase Corpus (MRPC) [45], which is composed of pairs of sentences extracted from news sources on the web. Each sentence pair has been annotated by humans to determine whether they convey a paraphrase or semantic equivalence relationship. SentEval also includes datasets to evaluate models on semantic textual similarity, comprising pairs of sentences collected from news articles, forum discussions, news conversations, headlines, and image

and video descriptions labeled with a similarity score between 0 and 5. The goal is to evaluate how the cosine distance between two sentences correlates with a human-labeled similarity score through Pearson and Spearman correlations. Pearson correlation coefficient is a measure of the linear relationship between two variables. Its values range from -1 to 1, where 1 indicates a perfect positive linear relation, -1 value indicates a negative linear relation, and zero indicates no linear relationships. Pearson correction coefficient(r) is calculated using the following formula [46]:

$$r = \frac{\Sigma \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\left[\Sigma \left(X_i - \bar{X}\right)^2 \Sigma \left(Y_i - \bar{Y}\right)^2\right]^{1/2}} \tag{8}$$

where X_i and Y_i are the data points for variables X and Y respectively, and $\bar{X}$ and $\bar{Y}$ denote the means of variables X and Y.

Spearman's correlation measures the degree and direction of association between two variables based on their ranked values. It evaluates the strength of the relationship between two variables using a monotonic function. Spearman's correlation can be calculated using the following formula [47]:

$$\rho = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)} \tag{9}$$

where d_i is the difference in two ranks of each observation and n is the number of observations. SentEval reports both the average and weighted average (by the number of samples in each subtask) of the Pearson and Spearman correlations.

For the semantic similarity experiment, we used twelve datasets including WS-353-All, Mturk-77, WS-353-REL, YP-130 [48], SimLex-999 [49], WS353 [50], and Men [51,52]. The cosine similarity between corresponding vectors was calculated, and the detailed performance against each of the twelve datasets is reported in Table 1. We see consistent and significant improvement from the three embedding models across all twelve datasets with an average improvement of 8 percent. This validates the incorporation of external knowledge from WordNet when our proposed autoencoder model and variance-invariance constraints are used.

Table 2. Average purity measure of the original embeddings and our proposed model embeddings on a concept categorization task

Model	Ap	BLESS	Battig	ESSLI_2c	ESSLI_2b	ESSLI_1a
fastText-original	65.17	75.55	43.47	60.00	**75.00**	86.36
ours	**68.90**	**76.23**	**44.73**	**68.88**	72.50	**87.30**
Word2Vec-original	62.68	69.50	**42.03**	64.44	82.22	79.55
ours	**64.13**	**71.50**	41.86	**68.00**	**85.39**	**81.81**
GloVe-original	63.68	82.34	**41.98**	**64.44**	**82.50**	75.84
ours	**65.18**	**84.18**	41.31	62.22	82.40	**79.27**

4.4 Concept Categorization

The concept categorization task is commonly used to evaluate the quality of word representations by grouping words into distinct categories. A purity measure is used to evaluate the quality of these clusters. In our experiment, we used six datasets, including AP [53], Bliss [54], [55], ESSLI_a, ESSLI_b, and ESSLI_c [56]). The AP dataset consists of 402 words grouped into 21 concept categories, while BLESS has 200 concepts classified into 17 semantic classes. Battig dataset includes 5231 words categorized into 56 taxonomic categories. ESSLI_a has 44 concrete nouns that fall under six semantic categories, while ESSLI_b features 40 nouns classified into three classes. ESSLI_c has 45 verbs divided into nine semantic classes. Table 2 shows the performance of different embedding models in terms of purity measure. Our proposed model has shown better performance than the original embedding of the three main embedding models. For fastText, our proposed model has outperformed the original model in 5 out of 6 datasets, with an average improvement of 3.3%. Likewise, our proposed model has outperformed the original Word2Vec model with an average improvement of 2% and an improvement in 5 out of 6 datasets. In the case of Glove, we have observed improvement as well, with an average of 1% across all the datasets. To summarize, our proposed model has outperformed baseline models, with an average improvement of 2% for purity measures. This indicates that our proposed model improves the performance of the concept categorization task (Table 3).

4.5 Ablation Study

We conducted an ablation study to assess the impact of variance in the loss function on model performance. We tested the model with different weights for the variance component in the loss function, using values of 0, 0.5, 1, 2, and 10, while keeping the weight of the invariance component constant at 2. We compared the performance of these models to baseline models. The experimental results show that our proposed model outperforms the baseline model by an average of 5.7% across all embedding models and datasets, even without the variance component. Adding the variance component (with a weight greater than zero) leads to a significant improvement, boosting model performance by 8% across different embedding models and datasets. However, we also observed a performance decrease when the variance weight was increased beyond 2. Based on these results, we recommend setting the variance weight to either 1 or 2 as

Table 3. Average Spearman's correlation results for our proposed model, using fastText word embeddings with varied weights assigned to the variance component in the loss function, on a semantic similarity task.

Model	WS -353 -ALL	MTurk -771	WS -353 -REL	YP -130	RW -STAN -FORD	WS -353 -SIM	VE -RB -143	MTurk -287	SIM -LEX -999	SIM -MC -30	RG -65	MEN -TR -3k
fastText-org	70.84	71.01	65.00	51.87	52.33	81.02	46.27	70.50	45.00	83.63	84.51	79.06
Method (variance = 0)	76.39	75.21	73.89	61.69	55.94	81.97	46.90	71.25	50.22	84.23	84.51	82.86
Method(variance = 0.5)	76.29	76.67	73.08	69.10	57.66	82.32	48.91	73.76	53.06	86.77	86.20	84.21
Method (variance = 1)	**76.69**	**76.84**	**73.31**	**70.48**	**57.86**	**82.90**	**49.02**	**74.03**	**52.90**	**87.48**	**86.83**	**84.08**
Method (variance = 2)	74.77	77.03	71.42	71.28	57.33	82.57	49.39	74.35	52.06	88.77	86.69	83.67
Method (variance = 10)	68.55	74.27	62.90	63.01	55.22	78.05	78.05	74.20	48.67	82.45	86.28	81.38

the optimal value in the loss function. Additionally, we set the variance of all dimensions of the word embeddings to 0.1, which yielded the best performance in our experiments. Increasing the variance beyond this value led to a decline in performance.

5 Conclusion

In this work, we introduced a new method to incorporate external knowledge extracted from lexical resources such as WordNet. Our approach is based on a post-processing technique which is an effective method for improving word embedding that does not require training from scratch. We utilize autoencoder as the base architecture for our model with variance and invariance terms in the loss function to control the quality of the processed word vectors. As a result, our method performs significantly better than the original word embedding models (Word2Vec, GloVe, and fastText) on semantic similarity tasks. In the future, we would like to include more knowledge into word embeddings using other lexical resources such as BabelNet and ConceptNet. We also intend to apply our method to other languages and to multi-lingual embeddings.

Limitations

We presented a method to improve the embedding space by utilizing external knowledge resources. Our method trains the model to minimize the distance between word pairs that share a specific type of relation extracted from WordNet. However, there are limitations to our approach. We only consider a predefined set of relationships and do not incorporate all available relations in WordNet. Additionally, we only include words that are present in the pre-trained embedding model, eliminating words consisting of multiple words or absent from the model from the training set. Our model does not support opposite relations, such as antonyms, as these require maximizing the distance between the word pairs. Furthermore, it is important to note that our model has primarily been tested in English. Further studies are needed to validate the generalizability of our model across different languages.

Ethics Statement. Our proposed model aims to improve the quality of pre-trained word embedding such as Word2vec, GloVe, and fastText. These models exhibit biases, such as gender or cultural biases, which could lead to fairness issues or unintended discriminatory outcome when applied in to real word application. Additionally, incorporating the WordNet dataset in our model could introduce more biases. Therefore, it is critical to exercise caution when utilizing these embeddings in sensitive domains domains, such as finance, healthcare, or legal contexts where the impact is significant.

References

1. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1006–1011 (2015)
2. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. In: Transactions of the Association for Computational Linguistics (2015)
3. Schakel, A.M., Wilson, B.J.: Measuring word significance using distributed representations of words. arXiv preprint arXiv:1508.02297 (2015)
4. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166 (2014)
5. Glavaš, G., Vulić, I.: Explicit retrofitting of distributional word vectors (2018)
6. Albujasim, Z., Inkpen, D., Guo, Y.: Word embedding interpretation using co-clustering. In: CS & IT Conference Proceedings, vol. 12, no. 22 (2022)
7. Rajaee, S., Pilehvar, M.T.: A cluster-based approach for improving isotropy in contextual embedding space. arXiv preprint arXiv:2106.01183 (2021)
8. Kaneko, M., Bollegala, D.: Autoencoding improves pre-trained word embeddings. In: International Committee on Computational Linguistics (2020)
9. Albujasim, Z., Inkpen, D., Han, X., Guo, Y.: Improving word embedding using variational dropout. In: The International FLAIRS Conference Proceedings, vol. 36 (2023)
10. Mrkšić, N., et al.: Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. Trans. Assoc. Comput. Linguist. **5**, 309–324 (2017)
11. Vulić, I., Glavaš, G., Mrkšić, N., Korhonen, A.: Post-specialisation: retrofitting vectors of words unseen in lexical resources. arXiv preprint arXiv:1805.03228 (2018)
12. Tissier, J., Gravier, C., Habrard, A.: Dict2vec: learning word embeddings using lexical dictionaries. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 254–263 (2017)
13. Mrkšić, N., et al.: Counter-fitting word vectors to linguistic constraints. arXiv preprint arXiv:1603.00892 (2016)
14. Roy, A., Pan, S.: Incorporating extra knowledge to enhance word embedding. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 4929–4935 (2021)
15. Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., Hu, Y.: Learning semantic word embeddings based on ordinal knowledge constraints. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1501–1511 (2015)
16. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
17. Mu, J., Bhat, S., Viswanath, P.: All-but-the-top: simple and effective postprocessing for word representations. arXiv preprint arXiv:1702.01417 (2017)
18. Vulić, I., Mrkšić, N.: Specialising word vectors for lexical entailment. arXiv preprint arXiv:1710.06371 (2017)
19. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: From paraphrase database to compositional paraphrase model and back. Trans. Assoc. Comput. Linguist. **3**, 345–358 (2015)

20. Rothe, S., Schütze, H.: Autoextend: extending word embeddings to embeddings for synsets and lexemes. arXiv preprint arXiv:1507.01127 (2015)
21. Nguyen, K.A., Köper, M., Walde, S.S.i., Vu, N.T.: Hierarchical embeddings for hypernymy detection and directionality. arXiv preprint arXiv:1707.07273 (2017)
22. Mrkšić, N., Séaghdha, D.O., Wen, T.-H., Thomson, B., Young, S.: Neural belief tracker: data-driven dialogue state tracking. arXiv preprint arXiv:1606.03777 (2016)
23. Bosc, T., Vincent, P.: Auto-encoding dictionary definitions into consistent word embeddings. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1522–1532 (2018)
24. Grill, J.-B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 271–284 (2020)
25. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
26. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 750–758 (2021)
27. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
28. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:1702.03859 (2017)
29. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
30. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (2014)
31. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
32. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
33. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
34. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
35. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1352–1362 (2013)
36. Conneau, A., Kiela, D.: Senteval: an evaluation toolkit for universal sentence representations. arXiv preprint arXiv:1803.05449 (2018)
37. Wang, S.I., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 90–94 (2012)
38. Li, X., Roth, D.: Learning question classifiers. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)

39. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: Semeval-2012 task 6: a pilot on semantic textual similarity. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and The Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 385–393 (2012)
40. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: * sem 2013 shared task: semantic textual similarity. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and The Shared Task: Semantic Textual Similarity, pp. 32–43 (2013)
41. Agirre, E., et al.: Semeval-2014 task 10: multilingual semantic textual similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 81–91 (2014)
42. Agirre, E., 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)
43. Agirre, E., et al.: Semeval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16–17; San Diego, CA, pp. 497–511. ACL, Stroudsburg (PA) (2016)
44. Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., Zamparelli, R.: Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. Lang. Resour. Eval. **50**, 95–124 (2016)
45. Dolan, B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Third International Workshop on Paraphrasing (IWP2005) (2005)
46. Rodgers, J.L., Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. Am. Stat. **42**(1), 59–66 (1988)
47. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **100**(3/4), 441–471 (1987)
48. Yang, D., Powers, D.M.: Verb Similarity on the Taxonomy of WordNet. Masaryk University (2006)
49. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: evaluating semantic models with (genuine) similarity estimation. Comput. Linguist. **41**(4), 665–695 (2015)
50. Finkelstein, L., et al.: Placing search in context: the concept revisited. In: Proceedings of the 10th International Conference on World Wide Web, pp. 406–414 (2001)
51. Bruni, E., Tran, N.-K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res. **49**, 1–47 (2014)
52. Luong, M.-T., Socher, R., Manning, C.D.: Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp. 104–113 (2013)
53. Poesio, M., Almuhareb, A.: Identifying concept attributes using a classifier. In: Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, pp. 18–27 (2005)
54. Baroni, M., Lenci, A.: How we blessed distributional semantic evaluation. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics (2011)
55. Baroni, M., Lenci, A.: Distributional memory: a general framework for corpus-based semantics. Comput. Linguist. (2010)
56. Almuhareb, A.: Attributes in lexical acquisition. Ph. D. dissertation, University of Essex (2006)

Multimodal Sign Language Model

Christian Velasquez Borasino(✉), Giorgio Mancusi Barreda, and Rody Vilchez Marin

Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru
{u202218075,U202216613,U202216562}@upc.edu.pe

Abstract. Sign language translation (SLT) plays a key role in improving accessibility for deaf communities, but faces major challenges such as limited parallel video-text data. In this work, we present Imitator, an architecture that learns to produce text embeddings that align with those of a pretrained large language model (LLM), such as LLaMA. Our approach combines lightweight 1D convolutions and a Transformer encoder with a cross-attention mechanism using learnable token queries to flexibly map variable-length video sequences into fixed-length text representations. The preliminary results show promising alignment metrics, suggesting the potential of this imitation-based strategy for SLT. Imitator model achieves a mean squared error with cosine similarity (MSE + CosSim) of 8×10^{-4} on the validation set. These results highlight the potential of imitation-based embedding alignment as a lightweight alternative for sign language translation in low-resource contexts.

Keywords: Sign language · embeddings · distillation · LSP

1 Introduction

Sign language is the main mode of communication and expression for people who are deaf or hard of hearing [10]. Despite its importance, a significant communication gap persists between sign language users and non-users [7], further exacerbated by the lack of accessible technological tools specifically designed for this population. Bridging this gap requires the development of effective translation systems capable of interpreting sign language accurately and robustly [4].

Sign language consists of a wide variety of complex hand movements, facial expressions, and body gestures, making automatic translation a highly challenging task [3]. One of the major obstacles for Peruvian Sign Language (LSP, lengua de señas peruana) translation is the scarcity of large-scale datasets containing parallel sign videos and textual annotations [7]. This data limitation makes it difficult to train end-to-end models, which often rely on direct mapping from sign-language input to textual output [10].

Traditional SLT approaches typically involve predicting intermediate gloss sequences as a surrogate representation before generating text [4]. However, these models tend to suffer from limited accuracy and poor generalization, especially in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 327–341, 2026.
https://doi.org/10.1007/978-3-032-20322-9_23

low-resource settings [3]. Recent studies confirm that even state-of-the-art gloss-based methods still outperform direct gloss-free approaches, although the gap is slowly narrowing thanks to advances in pretraining and contrastive learning [3,13].

In this work, we propose *Imitator*, an architecture that learns to imitate the text representations of a large language model (LLM) such as LLaMA [1], bypassing the need for intermediate gloss labels while maintaining semantic alignment with the target text.

2 Related Work

Liu et al. [4] present SCOPE, a context-oriented framework for sign language recognition (SLR) and translation (SLT) tasks. This approach employs a multimodal encoder that integrates keypoints with textual embeddings extracted from a large language model (LLM). For translation, the language model is fine-tuned using prior conversational information. The authors also introduce a new dataset containing 72 h of Chinese Sign Language videos in dialog contexts. SCOPE outperforms previous models on the PHOENIX-2014T and CSL-Daily benchmarks, demonstrating significant improvements in context-aware text generation.

Huang et al. [4] propose SignCL, a contrastive learning framework for gloss-free SLT. Its objective is to reduce the performance gap between gloss-free and gloss-based methods by enforcing alignment between visual features and target text embeddings through contrastive loss. Despite these advances, gloss-based models still outperform SignCL on datasets such as CSL-Daily and RWTH-PHOENIX-Weather.

Li et al. [3] introduce ADTR-VLP, a pretraining framework for gloss-free SLT that combines visual–language contrastive objectives with a masked modeling task. This method achieves competitive performance with fewer parameters and narrows the performance gap with gloss-based approaches. However, the translation quality remains below that of fully supervised gloss-based systems.

Zhao et al. [13] extend LLaVA to the sign language domain by leveraging a vision-language model pretrained on sign videos and aligning visual inputs with LLM text embeddings. This approach further reduces the gap between gloss-based and gloss-free systems, demonstrating that large multimodal models can be adapted to sign language without explicit gloss supervision.

Taken together, these works illustrate a clear shift toward gloss-free translation by leveraging contrastive objectives, masked modeling, and multimodal alignment with large language models. Nevertheless, most existing approaches still depend on expensive pretraining or limited datasets, leaving an open opportunity for lightweight architectures capable of directly aligning visual inputs with LLM latent spaces. Our work addresses this gap by imitating LLM embeddings without gloss supervision, improving portability and reducing annotation costs.

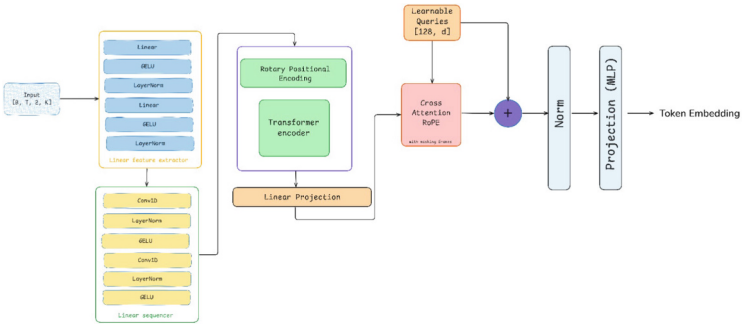


Fig. 1. High-level architecture of the Imitator model including feature extraction, temporal encoding, and embedding alignment with LLaMA.

3 Methodology

3.1 The Imitator Architecture

The Imitator architecture is designed to map sequences of human keypoints into fixed-length text embeddings that align with a target language model (e.g., LLaMA), as illustrated in Fig. 1. The model receives a temporal sequence of 2D skeleton keypoints per video frame and encodes them into compact embeddings through a convolutional-attentive pipeline, mixing convolutional networks and transformers.

Feature Projection. A multi-layer perceptron (MLP) reduces the input dimensionality and performs early feature extraction. The MLP consists of two linear projections with GELU activations and LayerNorm, reducing from D to $hidden/2$,

Temporal Modeling. The reduced features are passed through a lightweight temporal encoder composed of two 1D convolutions. The first convolution (kernel size $= 3$, stray $= 1$ and padding $= 1$) extracts local temporal features; the second (kernel size $= 1$) enables channel mixing. The outputs are normalized and reprojected to the transformer dimension $hidden = 512$.

An *adaptive average pooling* layer is applied to reduce the temporal dimension from T to T' (by default $T' \approx 2L$), while the frame-level padding mask is propagated through *adaptive max pooling* over the binary validity map, prior to the Transformer encoder.

Rotary Positional Embeddings (RoPE) Instead of the classical sinusoidal positional encoding proposed by [12], we adopt rotary positional embeddings (RoPE) [11] as a mechanism to encode temporal information. Unlike fixed absolute encodings, RoPE incorporates the position via learned rotations in the embedding space, allowing the model to reason directly over relative distances between tokens. This is particularly beneficial for sign language, which exhibits natural cyclic or repetitive patterns at both the word and sentence level. Moreover, RoPE encodes positional noise, such as jitter in keypoint detection, as

small, smooth rotations in the query and key vectors, leading to stable attention weights and improved robustness to temporal variation. In this work, we implement a RoPE transformer encoder modifying current implementation of transformer encoder layer by PyTorch. Our implementation is based on the modification of a `multi_head_attention_forward` function to implement a RoPE operation into a MultiHeadAttention module.

To capture long-range dependencies across frames, we employ a transformer encoder with rotary positional embeddings (RoPE), as previously described. The encoder consists of $n = 2$ layers, each with six attention heads, GELU activations, and dropout. Layer normalization is applied in pre-norm. The PyTorch transformer encoder layer was re-implemented to include a new argument, `use_rotary=True`, enabling RoPE with a default base of `1000`.

Projection to Embedding Space. Each time step is independently projected into a target output space of dimension 3072. We selected an output dimensionality of 3072 to exactly match the token embedding width of the LLaMA-3.2B model. This ensures one-to-one compatibility between the visual representation and the latent text space of the LLM, enabling direct cosine-based supervision without additional projection layers. Target embeddings are obtained by passing the tokenized transcript through the LLaMA embedding layer, using Hugging Face's Transformers library to extract the embeddings corresponding to each input token. The resulting tensor [B, T', `output_dim`] is interpreted as a set of visual keys and values.

Cross-attention Head. Motivated by the need to align video sequences with textual outputs of up to 128 tokens the maximum number of tokens observed after tokenizing the transcript using the LLaMA model's tokenizer we define a fixed set of $L = 128$ learnable token queries with positional encoding to ensure that the tokens remain distinguishable based on the temporal dimension of the video. These act as latent slots, each extracting semantically meaningful information from the variable-length visual input via multi-head cross-attention. Each query $Q \in \mathbb{R}^{B\times L\times d}$ attends over the frame-level representations $M \in \mathbb{R}^{B\times T\times d}$, using sixteen attention heads and a dropout rate of 0.2. This mechanism enables the model to flexibly pool visual information into a structured output space of fixed length, suitable for alignment with language model embeddings.

The cross-attention block outputs a structured sequence of token-level embeddings rather than a single global vector. Each learnable query acts as a pseudo-token that aggregates information from different temporal regions of the video, analogous to language tokens in a sentence. During training, these query outputs are aligned position-wise with the corresponding LLaMA token embeddings. In downstream applications, the entire set of 128 aligned vectors can be averaged or pooled to form a global representation when required.

Inspired by Perceiver IO [1], the attended output is added residually to the input queries and normalized:

$$X = \text{LayerNorm}(Q + \text{MultiHeadAttn}(Q, M, M)).$$

To improve the representational capacity and training stability, the result is passed through a residual feedforward block composed of two linear layers (with expansion factor 2), GELU activation, and 0.1 dropout—analogous to the latent update steps in Perceiver IO.

3.2 Loss Function

To train the Imitator model, we employ a masked loss that combines a mean squared error (MSE) term and a cosine similarity term, encouraging alignment between the predicted token embeddings and the corresponding target embeddings produced by a frozen language model. The loss is computed at the token level, aligning each of the L predicted visual token embeddings with their textual counterparts from the LLaMA model.

A binary mask is used to ignore padding tokens in shorter sequences. Specifically, given predicted embeddings $\hat{E} \in \mathbb{R}^{B\times L\times d}$ and target embeddings $E \in \mathbb{R}^{B\times L\times d}$, we first truncate both to the shortest common length ($L^\star$) to prevent shape mismatches. Let $v \in \{0,1\}^{B\times L^\star}$ denote the validity mask, where $v_{b\ell} = 1$ for valid tokens and $v_{b\ell} = 0$ for padding.

The masked mean squared error term is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{\sum_{b=1}^{B}\sum_{\ell=1}^{L^\star} v_{b\ell}} \sum_{b=1}^{B}\sum_{\ell=1}^{L^\star} v_{b\ell} \cdot \frac{1}{d} \left\| \hat{E}_{b\ell} - E_{b\ell} \right\|_2^2.$$

Similarly, the masked cosine similarity loss is computed as:

$$\mathcal{L}_{\cos} = \frac{1}{\sum_{b=1}^{B}\sum_{\ell=1}^{L^\star} v_{b\ell}} \sum_{b=1}^{B}\sum_{\ell=1}^{L^\star} v_{b\ell} \cdot \left(1 - \frac{\hat{E}_{b\ell}}{\|\hat{E}_{b\ell}\|_2} \cdot \frac{E_{b\ell}}{\|E_{b\ell}\|_2}\right).$$

To balance the two objectives, we introduce a dynamic scaling factor:

$$\beta = \frac{\mathcal{L}_{\text{MSE}}}{\mathcal{L}_{\cos} + \varepsilon}, \quad \text{where } \varepsilon = 10^{-6}.$$

Gradients are not propagated through β.

The final total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \beta\, \mathcal{L}_{\cos}.$$

This formulation ensures that the model focuses exclusively on semantically valid tokens while dynamically adjusting the contribution of cosine similarity based on the relative magnitudes of the two losses. It thus enables stable and robust training even when sequence lengths and embedding magnitudes vary across samples. Importantly, the loss values are computed per token, reflecting the mean squared error over an embedding dimension of 3072.

This setup allows for robust learning even when the target embedding lengths vary across samples (Fig. 2).

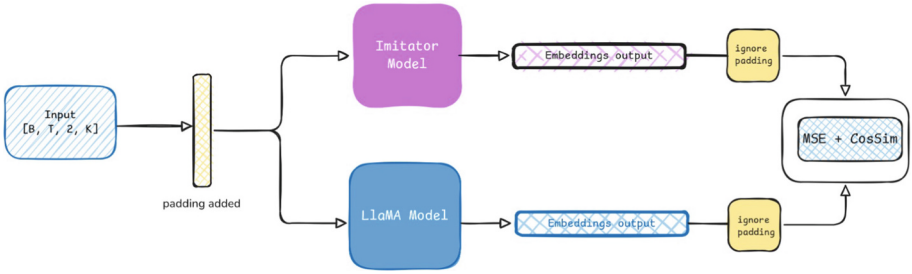


Fig. 2. Imitator Loss function

3.3 Optimization Schedule

We adopt a learning rate schedule based on a linear warm-up followed by cosine decay, a strategy shown to stabilize training and improve generalization in transformer-based models [5,12]. Specifically, the learning rate η_t at step t is defined by:

$$\eta_t = \begin{cases} \frac{t}{T_w}, & \text{if } t < T_w \\ \frac{1}{2}\left(1 + \cos\left(\pi \cdot \frac{t - T_w}{T - T_w}\right)\right), & \text{otherwise} \end{cases}$$

where T_w is the number of warm-up steps (typically 5 epochs), and T is the total number of training steps. The scheduler is implemented via PyTorch's `LambdaLR`, which dynamically adjusts the learning rate each step. This approach allows the optimizer to ramp up gradually at the start, avoiding instability, and then anneal smoothly to improve convergence as training progresses.

3.4 Data Preprocessing and Preparation

Video Selection. To improve data quality, we excluded videos whose labels contained special characters (such as symbols). After filtering, a total of 4,100 videos were retained for training and validation.

Keypoint Extraction. To obtain a standardized representation across all sign language videos, we used the RTMPose-L model from the RTMlib repository [2]. This model extracts 133 keypoints per frame, including body, hand, and facial landmarks. This 2D representation captures the spatial complexity of sign language gestures while remaining compact enough for efficient processing. All videos were processed at 30 fps, and keypoints were extracted at the original video resolution.

Normalization and Padding Handling. To standardize spatial scale and position, keypoints are normalized using a shoulder-centered coordinate system. The skeleton is centered at the midpoint between the shoulders and scaled by

the shoulder distance. Keypoints with null coordinates (e.g., (0,0)) are masked and replaced with out-of-range values (e.g., (-3,-3)), preserving their presence without affecting model training.

Filtering Unstable Keypoints. To reduce noise and avoid overfitting, we compute the temporal variance of each joint and retain the N = 89 most stable keypoints, based on empirical evaluation. Less stable joints are set to zero out across the sequence, simplifying the input and enhancing model robustness.

Removed Extra Keypoints. To reduce unnecessary keypoints, such as those corresponding to legs and face border, we decreased the total number from 133 to 89. As shown below, some of the removed keypoints include those on the legs and along the face border (Fig. 3).

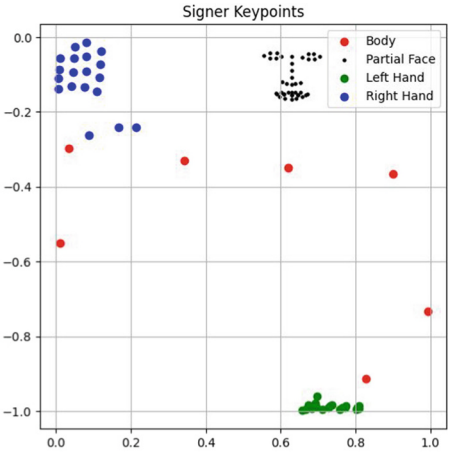


Fig. 3. Signer Keypoints

Embedding Generation and Target Construction. For each video, we generate a fixed-size (3072-dimensional) textual embedding using the LLaMA-3.2B model. These embeddings serve as soft targets during training, enabling alignment in shared embeddin space without the need for gloss annotations.

Unlike methods such as SignCL or LLaVA-SLT, our architecture does not depend on retraining the LLM when it is updated. Imitator remains compatible as long as embedding space itself is unchanged. This design allows the learned embeddings to be directly indexed within the LLM's latent space, thereby eliminating retraining costs and enhancing portability across models.

HDF5 Storage Format. All preprocessed data (keypoints, embeddings, labels) are stored in HDF5 format using a hierarchical structure. This format supports modular dataset access, fast loading, and version control. Each dataset is organized into separate groups, enabling selective activation during training or ablation studies.

3.5 Data Loading and Sampling

An overview of the full data flow—from keypoint input to embedding alignment—is shown in Fig. 4. This diagram summarizes how each component processes the input to produce LLM-aligned embeddings.

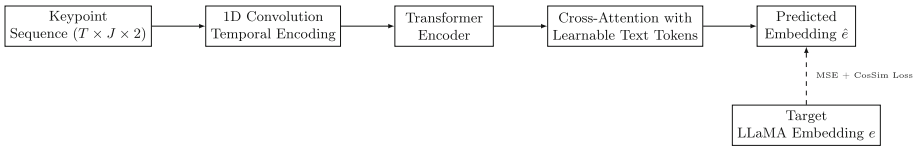


Fig. 4. Overview of the Imitator architecture. The input sequence of keypoints is processed via 1D convolutions and a Transformer encoder. A cross-attention module with learnable text tokens maps the temporal features to a fixed-dimensional embedding aligned with the LLaMA target.

4 Dataset and Experimental Setup

4.1 Datasets

LSA64. We evaluate our framework on LSA64 [9], an isolated Argentinian Sign Language dataset containing 3,200 video recordings of 64 lexical signs performed by 10 subjects (five repetitions per sign) under controlled conditions, with signers wearing colored gloves to facilitate hand tracking. Each sample consists of an RGB video sequence, from which we extract 2D keypoints using an off-the-shelf pose estimator. Our pipeline does not employ gloss annotations.

Test Videos for the Peruvian Sign Language Dictionary ML Model. Additionally, the test video dataset for the Peruvian Sign Language Dictionary ML model [6], comprising 38 video samples (total size 78.1 MB) extracted from the LSP PUCP 305 dataset. These videos, published on May 22, 2024, are used to assess the dictionary's AI model accuracy; each sample is provided as an MP4 file with accompanying metadata.

Peruvian Sign Language (LSP) PUCP-DGI156. We further leverage the PUCP-DGI156 dataset [7], which offers original and sign-segmented MP4 videos, SRT subtitle files, and MediaPipe-based pose estimations (27 keypoints) of continuous Peruvian Sign Language. The dataset includes over 1 GB of segmented footage and 690 MB of pose data, segmented using ELAN software.

Peruvian Sign Language (LSP) PUCP 305. We further utilize the PUCP 305 dataset [8], which provides original and sign-segmented MP4 videos of continuous Peruvian Sign Language. The dataset comprises over 2 GB of video material, containing 526 recordings featuring various signers and signs.
Although the primary training data originate from the Argentinian Sign Language (LSA), its structural and visual properties differ markedly from those of the Peruvian Sign Language (LSP), including differences in vocabulary, grammatical organization, and signer-specific conventions [7,10]. Such divergence introduces a cross-linguistic domain shift that provides a suitable scenario to evaluate the robustness and generalization capacity of the proposed gloss-free embedding model. To explicitly incorporate this variability into the learning process, both LSA and LSP datasets are merged prior to applying the data-splitting procedure detailed in Sect. 4.2.

4.2 Data Splits

All samples from the combined datasets are randomly shuffled without consideration of signer identity or gloss annotation. The partitioning is performed at the sample (video-fragment) level rather than by signer or gloss, yielding 3,280 samples (80%) for training and 820 samples (20%) for validation.The adopted intrasigner setting allows the same signer to appear in both partitions, which is consistent with our objective of assessing generalization across instances rather than across unseen signers. At this stage, no separate test set is defined; all quantitative evaluations are conducted on the validation set, as the present work constitutes an initial exploration of the proposed approach under cross-linguistic variability.

4.3 Cross Validation

We employ a 5-fold cross-validation protocol to improve robustness and reduce variance in reported metrics. All datasets are merged and randomly shuffled at the sample level, disregarding gloss and signer identity. In each fold, 80% of samples are allocated for training and 20% for validation. The reported metrics correspond to the average performance across all five folds. Note that this protocol is intra-signer: the same signer may appear in both splits. A held-out test set is not defined at this stage.

The detailed results for each fold are summarized in Table 1. We observe very low variability across folds: the training loss averages 6.84×10^{-4} with a standard deviation of 0.0005, while the validation loss remains consistently around 8.04×10^{-4}. This stability indicates robust training dynamics and consistent embedding alignment performance under the proposed cross-validation protocol. Importantly, these loss values are computed per token, reflecting the mean squared error over an embedding dimension of 3072.

Table 1. Training and validation losses per fold under the 5-fold cross-validation protocol (intra-signer). Mean and standard deviation (std) are reported across folds.

Fold	Train Loss	Val Loss
1	$6.44e^{-4}$	$8.04e^{-4}$
2	$8.51e^{-4}$	$8.16e^{-4}$
3	$6.53e^{-4}$	$8.06e^{-4}$
4	$6.30e^{-4}$	$8.10e^{-4}$
5	$6.42e^{-4}$	$7.71e^{-4}$
Average	$\mathbf{6.84e^{-4}}$	$\mathbf{8.01e^{-4}}$
Std	$\mathbf{9.37e^{-5}}$	$\mathbf{1.76e^{-5}}$

4.4 Training Scheme

The model was trained for 100 epochs using the AdamW optimizer with a base batch size of 64. Gradient accumulation across two mini-batches was employed, resulting in an effective batch size of 128 samples per optimization step. A linear warm-up followed by cosine decay was applied as the learning rate schedule. Training was performed on an NVIDIA L4 GPU with 24 GB of VRAM and required approximately 1 h to complete due to early stopping being enabled. The dataset was split into 80% for training and 20% for validation with cross-validation. Without signer-exclusive evaluation was conducted at this stage.

4.5 Model Metrics

Most of our training videos contain short utterances of only one to three words. In this regime, the visual evidence for each word spans a large portion of the sequence, which naturally yields high attention entropy: the cross-attention module behaves as a form of soft temporal pooling rather than a precise frame locator. This behavior is desirable—it prevents overfitting to a few frames and promotes robustness to signer speed variations, coarticulation, and pose-detection jitter. Based on the images below shows the model alignment and the Cross-Attention Map (Fig. 5).

The retrieval metrics (R@K) show a consistent improvement across epochs in both directions (video→text and text→video), approaching near-perfect alignment at higher K values. This suggests that the model progressively learns a shared embedding space where corresponding video and text samples become close neighbors. In the geometric metrics, the alignment MSE decreases while CKA remains high and the Procrustes error declines slightly, confirming better structural correspondence between the two modalities. Uniformity values for both visual and textual embeddings remain stable, suggesting that the model maintains a good distributional spread and avoids representation collapse.

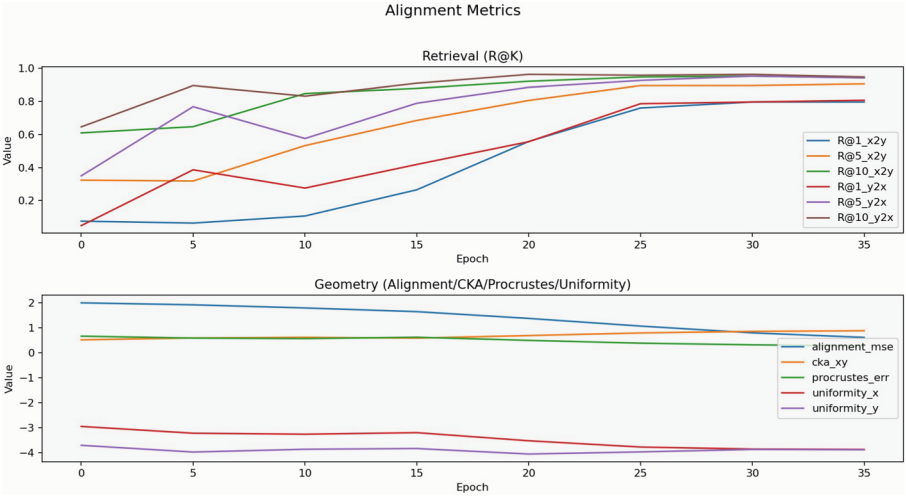


Fig. 5. Alignment metrics across training epochs.

Linguistic Evaluation. Although the current experiments focus on short utterances (one to three words), we also report a preliminary linguistic metric using BLEU scores computed with the `NLTK` implementation. Since most clips contain a single lexical item, BLEU-1 (unigram precision) provides the most meaningful signal, while higher-order scores (BLEU-2 and BLEU-4) naturally approach zero due to the absence of multi-word n-grams. Table 2 summarizes the results across three datasets: **Dataset 1** (LSA64) and **Datasets 3** and **5** (LSP). All values are reported on a corpus level. BLEU-1 for LSA64 reaches 0.5975 (59.8%), indicating that the model correctly associates approximately six out of ten visual tokens with their textual counterparts. Lower values on LSP (0.0244 and 0.0158) reflect the greater variability, signer diversity, and overall noise of that corpus. Because the evaluation clips contain isolated signs, these BLEU values should be interpreted as a proxy for word-level retrieval accuracy rather than full-sentence translation quality.

These results confirm that the model maintains consistent visual-textual associations even under a strictly gloss-free training regime. Here, BLEU-1 serves as a complementary measure to retrieval and alignment metrics, capturing the correctness of the lexical-level mapping learned by the model.
These results confirm that the model maintains consistent visual-textual associations even under a strictly gloss-free training regime. Here, BLEU-1 serves as a complementary measure to retrieval and alignment metrics, capturing the correctness of the lexical-level mapping learned by the model (Fig. 6).

The cross-attention map at epoch 45 reveals moderately structured token–frame associations, with a positive correlation ($\rho = 0.77$) between token index and attended frame position. The attention weights remain broadly distributed ($\mathcal{H} = 1.0$) and cover approximately 3.4% of the total frame area, indicating that

Table 2. Corpus-level BLEU scores for one-word evaluation clips.

Dataset	BLEU-1	BLEU-2	BLEU-4
Dataset 1 (LSA64)	0.5975	0.1543	0.0091
Dataset 3 (LSP)	0.0244	0.0079	0.0046
Dataset 5 (LSP)	0.0158	0.0020	0.0007

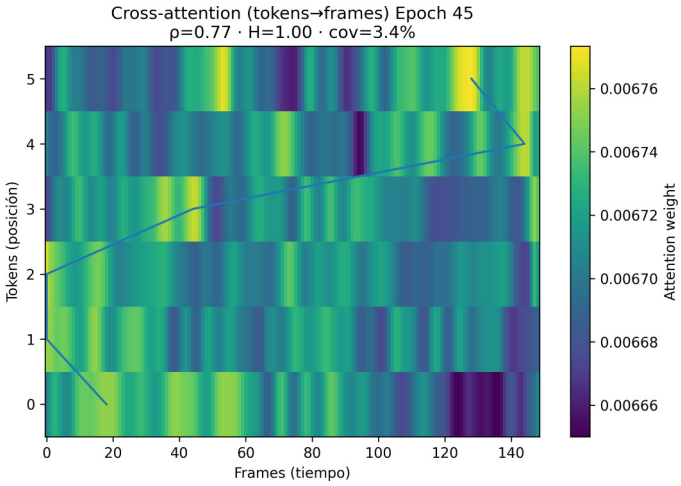


Fig. 6. Cross-attention visualization at epoch 45.

each token attends to multiple temporal regions rather than isolated moments. Given that the videos are short (1–3 words per clip), this broad attention is expected and reflects the compact temporal nature of the input rather than a failure of alignment.

5 Results

Under the 5-fold cross-validation protocol (Sect. 4.3), the Imitator model achieved an average validation loss of 8.01×10^{-4} and an average training loss of 6.84×10^{-4} (Table 1). These results are consistent with the single hold-out evaluation, where the model obtained a mean squared error with cosine similarity (MSE + CosSim) of approximately 8×10^{-4} on the validation set. This metric was computed between the predicted embedding, $\hat{e}$, and the target embedding, e, generated by the LLM.

Across all five folds, the model exhibited minimal variance, with training losses ranging from 6.30×10^{-4} to 8.51×10^{-4}, and validation losses remaining stable around 8.0×10^{-4}.

This consistency indicates robust convergence and suggests that the alignment mechanism generalizes effectively across different data partitions.

6 Discussion

The results confirm that Imitator is capable of extracting semantically meaningful representations from 2D keypoint sequences and aligning them with textual embeddings from a large language model. This demonstrates that rich linguistic information can be inferred directly from body motion, even in the absence of gloss annotations.

While the performance is encouraging, we acknowledge several limitations. First, the validation set may be overly optimistic due to the presence of the same signers in both training and validation (intrasigner overlap). Second, our results may reflect constraints imposed by consumer-grade hardware, such as suboptimal keypoint quality or limited batch sizes during training. Third, the dataset is derived from a single domain (news broadcast), which may affect generalization.

7 Conclusions and Future Work

7.1 Summary of Findings

In this work, we introduced Imitator, a novel sign language model that aligns sequences of 2D keypoints with text embeddings from a pretrained large language model (LLM) without relying on intermediate gloss annotations. The proposed architecture combines lightweight convolutional layers, Transformer encoders, and a cross-attention mechanism with learnable token queries to encode complex spatio-temporal dynamics of sign language into fixed-dimensional representations. By enabling direct supervision at the embedding level, Imitator facilitates faster adaptation to different sign languages or updated LLM backbones without requiring model retraining. Experimental results demonstrate that the model achieves high alignment accuracy using only motion-based inputs, validating the potential of gloss-free sign language translation in low-resource scenarios.

7.2 Extended Discussion and Future Directions

Imitator's alignment mechanism provides a promising foundation for low-cost and flexible sign language understanding systems. Unlike methods that require expensive retraining or fine-tuning of large language models, our approach aligns directly with the target embedding space, which improves adaptability to changes in the language model backend and reduces maintenance costs.

Despite the encouraging results, several avenues remain open for future exploration. A key direction is improving generalization, particularly by evaluating signer-independent scenarios and cross-dataset transfer to mitigate potential overfitting to the current corpus. Another critical aspect is downstream integration: assessing how the learned embeddings perform in tasks such as retrieval, classification, or LLM prompting, especially in multilingual or assistive contexts.

Further work is also needed in visualization and explainability, where tools to analyze embedding trajectories, token attention distributions, and interpretability maps could offer deeper insights into model behavior. From a benchmarking

perspective, incorporating established gloss-free SLT baselines would contextualize the effectiveness of our imitation-based strategy and highlight its comparative advantages.

In terms of applications, Imitator's lightweight design suggests potential for real-time deployment in sign recognition and communication assistance systems. External validation also remains an important milestone; defining a held-out test set and evaluating performance on additional public benchmarks will be essential to confirm real-world applicability. Extending the model toward continuous SLT tasks, where sequence-level textual targets are required, represents another promising challenge.

Finally, exploring LLM agnosticism by adapting the framework to alternative models such as Gemma-3B could provide further evidence of portability across architectures. Complementarily, advanced data augmentation techniques—including generative motion synthesis, signer-style transfer, or learned spatio-temporal perturbations—may enhance robustness and performance in low-resource scenarios.

Our findings underscore the feasibility of gloss-free learning through embedding alignment with pretrained LLMs. As accessibility tools continue to evolve, models like Imitator have the potential to play a central role in making digital systems more inclusive for deaf communities worldwide.

References

1. Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: general perception with iterative attention (2021). https://arxiv.org/abs/2103.03206
2. Jiang, T.: rtmlib. https://github.com/Tau-J/rtmlib (2023)
3. Li, J., et al.: ADTR-VLP: vision-language pretraining for gloss-free sign language translation. Mathematics **13**(6), 909 (2023)
4. Liu, Y., Zhang, W., Ren, S., Huang, C., Yu, J., Xu, L.: Scope: sign language contextual processing with embedding from LLMs (2024). https://arxiv.org/abs/2409.01073
5. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts (2017). https://arxiv.org/abs/1608.03983
6. Oporto, S.: Videos de prueba para el modelo de Machine Learning del Diccionario de Lengua de Señas Peruana (2024). https://hdl.handle.net/20.500.12534/FAHXZY
7. Rodriguez Mondoñedo, M.: Lengua de Señas Peruana (LSP) PUCP-DGI156 (2022). https://hdl.handle.net/20.500.12534/OJYYYS
8. Rodríguez-Mondoñedo, M., et al.: Lengua de Señas Peruana (LSP) - PUCP 305 (glosas) (2024). https://hdl.handle.net/20.500.12534/JU4OLG
9. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: a dataset of argentinian sign language. XX II Congreso Argentino de Ciencias de la Computación (CACIC) (2016)
10. Ronchetti, F., Quiroga, F.M., Estrebou, C., Lanzarini, L.: Handshape recognition for argentinian sign language using probsom (2023). https://arxiv.org/abs/2310.17427

11. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: enhanced transformer with rotary position embedding (2023). https://arxiv.org/abs/2104.09864
12. Vaswani, A., et al.: Attention is all you need (2023). https://arxiv.org/abs/1706.03762
13. Zhao, L., Xu, W., Li, X., He, L., Wu, Z.: LLAVA-SLT: towards large multimodal models for sign language translation. arXiv preprint arXiv:2412.16524 (2024)

Implementation of Intelligent User Interfaces and Their Impact on Software Quality: A Systematic Mapping Study

Pedro Aguilar-Encarnacion(✉), Carlos Iñiguez-Jarrín, and Julio Sandobalín

Escuela Politécnica Nacional, Av. Ladrón de Guevara E11-253, Quito, Ecuador
{pedro.aguilar,carlos.iniguez,julio.sandobalin}@epn.edu.ec

Abstract. Intelligent User Interfaces (IUIs) merge artificial intelligence and human–computer interaction to deliver context-aware, adaptive experiences. Yet research still lacks a cohesive view of how IUIs are integrated into software design/engineering and how they affect software quality. This systematic mapping study reviews 366 peer-reviewed works to identify gaps in IUI implementation across the development lifecycle. We analyze five dimensions: deployment environments, application domains, development phases, implementation techniques, and quality characteristics. IUIs are chiefly realized in web and mobile applications using machine learning, user modeling, and natural language processing. While gains in interaction and functional suitability are reported, security, reliability, and early-phase user involvement remain underexplored. Our findings call for stronger evaluation methods, standardized definitions, and tighter integration of AI-based interfaces into formal engineering workflows. The study offers a data-driven overview that bridges AI and software engineering, laying a foundation for quality-driven, automated adoption of IUIs.

Keywords: Artificial Intelligence · Intelligent User Interfaces · Quality Attributes · Software Engineering · Systematic Mapping Study

1 Motivation

The User Interface (UI) is the mechanism through which humans establish communication with machines [9]. UIs are essential in software applications due to their impact on user satisfaction regarding software usage. Furthermore, in the scientific domain, there is a dedicated Human–Computer Interaction (HCI) community that addresses challenges related to the analysis, design, implementation, and evaluation of UIs, aiming to optimize their functionality and reduce the complexity of interaction [12].

The growing demand for software applications, together with the wide variety of users, has created the need for context-sensitive UIs that meet the specific needs and preferences of users [25]. Intelligent User Interfaces (IUIs) have

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 342–357, 2026.
https://doi.org/10.1007/978-3-032-20322-9_24

emerged as a response to this demand. IUIs are capable of adapting and responding effectively to users' specific needs, aiming to improve the efficiency, effectiveness, and naturalness of interaction through user model representation, reasoning, and action [13,22]. In addition, IUIs employ concepts and techniques from Artificial Intelligence (AI), HCI, and Software Engineering (SE) [18].

The literature reports several secondary studies on IUIs, addressing application domains such as healthcare and telemedicine [28], the Internet of Things (IoT) and smart cities [11], evaluation methods [4], and research trends in the field [4,5]. In these studies, IUIs are interfaces that learn from user behavior and adapt the UI to context (e.g., interfaces that modify their layout and options).

In [28], the authors studied IUIs applied to Ambient Assisted Living technologies and concluded that their development is mainly concentrated in healthcare and telemedicine. They emphasize the need for research on multisensory interactions. [11] analyzed design trends for IUI in IoT-based systems, highlighting the need for new approaches to improve accessibility and usability. [5] analyzed IUIs and Adaptive User Interfaces (AUIs), revealing the lack of comparative studies that could guide efficient implementation of these interfaces. Lastly, [4] identified shortcomings in the definitions, metrics, and evaluation of IUIs, concluding that it is important to distinguish IUIs from AUIs and to establish clear evaluation criteria.

Although prior secondary reviews [4,5,11,28] show significant growth in IUI research, they also reveal limited interest in the implementation and evaluation of IUIs from a software development perspective. To the best of our knowledge, there is no secondary study examining the impact of IUIs on software quality, nor are there proposals for actionable strategies to support their adoption in software product development. Therefore, it is essential to conduct an up-to-date systematic mapping study (SMS) on IUIs focused on identifying and defining the challenges they pose to software quality and their integration within the software development process.

This study lies at the intersection of artificial intelligence applied to adaptive interfaces and software design and engineering processes, aiming to provide guidelines for their effective integration into quality software products.

The paper is organized as follows: Sect. 2 details the mapping method; Sect. 3 presents results (trends, approaches, gaps); Sect. 4 discusses findings, limitations, and implications; Sect. 5 concludes and outlines future work.

2 Research Method

This systematic mapping study follows the guidelines proposed by [6,20] and [19]. The study was conducted in three stages: *planning*, *conducting*, and *reporting*. The planning and conducting stages are described in this section, while the reporting stage is presented in the results section.

2.1 Planning Stage

In the planning stage, the following activities were carried out to establish the review protocol: definition of the research question, search for primary studies, inclusion and exclusion of articles, quality assessment, definition of the data extraction strategy, and selection of synthesis methods.

Research Question. The aim of this study is to identify existing research gaps in the development of IUIs, guided by the following research question: ***How do researchers and practitioners implement Intelligent User Interfaces in the software development process, and what is their effect on quality?***

We split the broad research question into five subquestions (Table 1). Q1–Q4 describe the implementation context; Q5 identifies the ISO/IEC 25010 quality characteristics that are integrated by the implementation of IUIs.

Table 1. Research Subquestions

Research Question	Motivation
Q1. In which types of production environments are IUIs implemented?	To identify the environments where IUIs are deployed.
Q2. In which types of applications are IUIs used and applied?	To determine the types of software in which IUIs are used and applied.
Q3. In which activities of the software development process is the implementation of IUIs addressed?	To explore which activities in the software development process consider the inclusion of IUIs.
Q4. What are the most used techniques in the development of IUIs?	To identify the techniques applied in the development of IUIs.
Q5. Which quality attributes are affected by the implementation of IUIs?	To identify the quality characteristics, according to ISO/IEC 25010, that are integrated or impacted by the implementation of IUIs.

Primary Studies Search. The search for primary studies was carried out in the following digital libraries: Scopus [10], Web of Science [8], IEEE Xplore [16], ACM Digital Library [3], and SpringerLink [29]. The search string was defined based on the key terms derived from the main research question, using the PEO (Population, Exposure, and Outcomes) framework [24], as shown in Table 2.

The present study adopted the broad approach recommended by [26] and did not include specific outcomes in the search string. Moreover, as [30] notes, in the field of IUI research, the term "intelligent" is not always used to describe the proposed solutions. Consequently, the search was expanded to include terms such as "adaptive" and "adaptable", as they are also commonly used in this context. This strategy allows for a comprehensive view of the research area, avoiding bias or limiting the scope of the analysis. Table 3 presents the search string used.

Table 2. Application of PEO to the Main Research Question

Research Question	Question Elements
How do researchers and practitioners implement IUIs in the software development process, and what is their effect on quality?	**Population:** Intelligent User Interfaces (IUIs) **Exposure:** Software Development **Outcome:** Implementation, Quality

Table 3. Keywords and Operators Used in the Search String

Term	***Keywords and Operators***
Intelligent User Interfaces	("intelligent interface*" OR "smart interface*" OR "intelligent user interface" OR "intelligent user interfaces" OR "IUI" OR "adaptive user interface" OR "adaptive user interfaces" OR "AUI" OR "adaptable user interfaces") AND
Software Development	("software" OR "software develop*" OR "software engineering")

The search was conducted within the metadata fields: title, abstract, and keywords. The syntax of the search string was adapted to ensure its correct application in each digital library. In the specific case of the Springer database, access restrictions prevented limiting the search to the defined metadata fields. Consequently, the search was performed across all available metadata in articles and conference proceedings with content available in preview mode. This approach enabled the evaluation of primary studies without compromising the integrity of the analysis, prioritizing relevant and accessible sources within the existing limitations.

To validate the search strategy, the results were compared against a sample of five selected primary studies [1,2,12,14,23], which are considered essential.

[7] highlight that a major milestone in the development of IUIs occurred in 1997, with the introduction of Microsoft's "Office Intelligent Assistant." Since then, advances in the IUI field have been limited, experiencing renewed interest in recent years driven by the technological boom in AI. Therefore, a review period was defined from 1997 to 2024.

Article Selection. Articles were included if they met at least one of the following criteria:

- Research articles presenting work on IUIs in the software development process and software quality.
- Research articles that present examples or empirical studies on implementing IUIs in software products, where IUIs are defined as interfaces that learn from user behavior and/or context and adapt accordingly (e.g., automatic UI generation; runtime adaptation/personalization; and related approaches with adaptive behavior).

Articles were excluded if they met at least one of the following criteria:

- Duplicate articles of the same study found in different databases.
- Introductory articles for special issues, books, or workshops.
- Articles not written in English or Spanish.
- Short articles with fewer than five pages.

The reliability of the include/exclude screening was assessed using Fleiss' Kappa. This measure is scored between 0 (poor agreement) and 1 (full agreement). Three independent raters classified a random sample of 30 candidate studies for inclusion in the mapping study. The resulting Fleiss' Kappa was **0.86**. This indicates an acceptable level of agreement among raters. [21].

Quality Assessment. We evaluated study quality with a three-point Likert questionnaire adapted from established Software Engineering secondary-study guidelines [6,19,20]. It comprised two subjective (SQ) and one objective (OQ) items assessing depth and relevance. Three reviewers rated all records independently; disagreements were resolved by discussion and, when needed, adjudication.

- **SQ:** *Does the study provide a clear and detailed analysis of IUIs in software development?*
- **SQ:** *Are the results of the study well-structured and do they provide clear and relevant information?*
- **OQ:** *Has the study been cited by other authors?*

The possible answers for the subjective questions were: "Agree (+1)", "Partially agree (0)", and "Disagree (-1)". For the objective question, the possible answers were: "Relevant (+1)" if the article had been cited by more than five authors, "Partially relevant (0)" if cited by one to five authors, and "Not relevant (-1)" if it had not been cited. For publications from the year 2024, a minimum score of "Partially relevant (0)" was assigned to avoid penalizing new articles that have not yet had time to accumulate citations.

After analyzing the articles, each reviewer assigned a score for each question. The total score ranged from -3 to +3. Articles with a final score of +2 or +3 were considered relevant. The quality assessment was conducted to identify studies that could contribute to answering each research subquestion, not to exclude any study from the systematic mapping.

Definition of the Data Extraction Strategy. A set of possible answers was established for each research subquestion in order to standardize the selection criteria.

For Q1, the environment in which the IUI operates was considered:

- **Local/Native:** The IUI runs directly on the user's device without relying on external infrastructure.
- **Cloud Computing:** The IUI operates through remote servers.
- **Hybrid:** The IUI combines local execution with remote server-based services.

For Q2, an article may be classified as:

- **Desktop application:** If the IUI is installed and runs on a personal or laptop computer.
- **Web application:** If the IUI operates through a browser and runs in an online environment.
- **Mobile application:** If the IUI runs on mobile devices such as smartphones or tablets.
- **Embedded application:** If the IUI runs directly on a specialized device or dedicated hardware, without relying on an external application.

For Q3, the classification was based on the structure of the software engineering process proposed by [27], which identifies five core activities. However, since primary studies may involve experiments such as case studies, this classification was adapted to the IUI development context. An article may mention one or more of the following activities:

- **Communication:** Describes interaction with users, clients, or experts to gather requirements and establish evaluation criteria for the IUI.
- **Planning:** Defines a plan for the development and implementation.
- **Modeling:** Defines prototypes or simulations of the IUI to analyze its design and functionality before final implementation.
- **Construction:** Implementation and integration of the IUI.
- **Deployment:** Delivers the IUI, including feedback collection, performance evaluation, and adjustments based on results.

For Q4, the term *technique* refers to the methods used to develop IUIs [15]. For data extraction, a systematic process was used involving *keywording*, or keyword detection, in order to structure a classification scheme reflecting the techniques employed. This process was carried out in the following stages:

- **Review:** Abstracts were analyzed to identify key terms related to IUI development techniques; if insufficient, the introduction or conclusion was reviewed.
- **Identification:** Based on the selected concepts, keywords representing the techniques used in IUI development were extracted.
- **Development of the Classification Scheme:** Keywords were grouped into categories representing the main IUI development techniques.
- **Scheme Update or Article Classification:** If additional techniques were identified in new studies, the classification scheme was updated.
- **Results:** The techniques used in IUI development were identified.

For Q5, the definitions of software quality established in the ISO/IEC 25010 standard [17] were used. An article may mention one or more quality characteristics: functional suitability, usability, compatibility, performance efficiency, reliability, flexibility, maintainability, safety and security.

Selection of Synthesis Methods. The study applied quantitative and qualitative synthesis. Quantitative analyses counted studies per subquestion, produced frequency graphs [6], and tallied articles by year. Qualitative synthesis summarized IUI benefits and limitations by subquestion.

2.2 Conducting Stage

Table 4 summarizes the review: 366 articles met the inclusion/exclusion criteria. Cross-database duplicates were consolidated—counting each study once (priority: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SpringerLink) removing 2045 records.

Table 4. Results of the Review Execution Stage

Source	Potential Articles	Selected Articles
Scopus	526	145
Web of Science	183	85
IEEE Xplore	182	22
ACM Digital Library	61	0
SpringerLink	1459	114
Total Search Results	**2411**	**366**

3 Results and Discussion

This section presents and discusses the results from the systematic mapping study. Table 5 summarizes, for each research subquestion and its possible answers, the number of reviewed articles. The results obtained for each subquestion are analyzed and discussed below.

Table 5. Systematic Mapping Study Results

Research Subquestion	Possible Answers	Number of Studies	%
Q1. Production environments where IUIs are implemented.	Local/Native	182	49.73
	Cloud Computing	123	33.61
	Hybrid	21	5.74
Q2. Types of software in which IUIs are used.	Desktop Application	100	27.32
	Web Application	120	32.79
	Mobile Application	60	16.39
	Embedded Application	92	25.14

continued

Table 5. continued

Research Subquestion	Possible Answers	Number of Studies	%
Q3. Software process activities where IUIs are considered.	Communication	141	38.52
	Planning	0	0
	Modeling	323	88.25
	Construction	313	85.52
	Deployment	203	55.46
Q4. Techniques used to develop IUIs.	Multimodal Interaction and Sensors	97	26.5
	Modeling and Ontologies	147	40.16
	Interface and UX Design	45	12.3
	Algorithms and Models	154	42.08
	Data and Signal Processing	181	49.45
	Evaluation and Metrics	14	3.83
	Rule-Based Systems	10	2.73
	Machine Learning and AI	184	50.27
	Integration and Development	42	11.48
	Others	4	1.09
Q5. Quality attributes impacted by the integration of IUIs.	Functional Suitability	288	78.69
	Interaction Capability	298	81.42
	Compatibility	1	0.27
	Performance Efficiency	279	76.23
	Reliability	114	31.15
	Flexibility	197	53.83
	Maintainability	134	36.61
	Security	37	10.11
	Safety	0	0

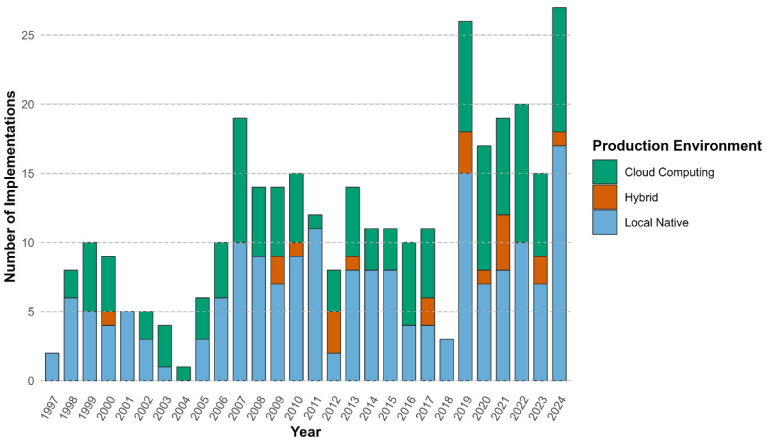


Fig. 1. Trend in the use of production environments for implementing IUIs between 1997 and 2024.

3.1 Production Environments in Which IUIs Are Implemented

For Q1, Fig. 1 shows the trend in the use of production environments for implementing IUIs between 1997 and 2024.

The local/native production environment is the most widely used (representing 49.73% of the reviewed articles), with significant peaks between 2007 and 2010 and a notable increase in 2024, indicating its current relevance. However, the Cloud Computing environment (33.61%) has become the alternative adopted in recent years, showing sustained growth from 2019 to the present. The hybrid environment (5.74%) maintains a limited presence compared to local/native and Cloud Computing environments, suggesting that the combination of local and cloud technologies has not yet been widely adopted. The year 2024 records the highest number of implementations across all environments, reflecting a growing trend in the adoption of IUIs.

On the other hand, 10.92% of the articles do not specify the production environment in which the IUIs are implemented.

In summary, the findings related to the production environments where IUIs are implemented are: i) since 2019, there has been a growing interest in adopting Cloud Computing as the environment for implementing web applications with IUIs, ii) the local/native environment remains predominant, especially in embedded and desktop applications, and iii) hybrid environments have not yet been widely adopted, suggesting technical challenges or a lack of standardization in integrating mixed infrastructures.

3.2 Types of Software that Use IUIs

For Q2, Fig. 2 shows that web applications exhibit progressive growth that accelerates from 2019 and reaches its highest point in 2024. Desktop applications predominate in the early years but begin to lose prominence after 2014.

The findings regarding the types of software in which IUIs are used can be summarized as:

i) web and mobile applications are the types of software that have experienced growing use of IUIs. This growth is driven by the demand for greater portability and personalization. ii) the use of IUIs in embedded applications has gained relevance in specialized domains, particularly in the fields of robotics and IoT. This reflects the expansion of IUIs into environments with specific and autonomous interactions, consolidating their presence in specialized devices.

3.3 Phases of the Software Process that Consider the Implementation of IUIs

For Q3, the results in Fig. 3 show that most articles refer to multiple phases of the software process (28.30% of articles analyze two phases, 36.89% three phases, and 23.22% four phases). A total of 6.28% of the articles refer to only one phase, and 4.64% of the articles do not consider any specific phase of the software process (cross-cutting works such as methodologies or best practices).

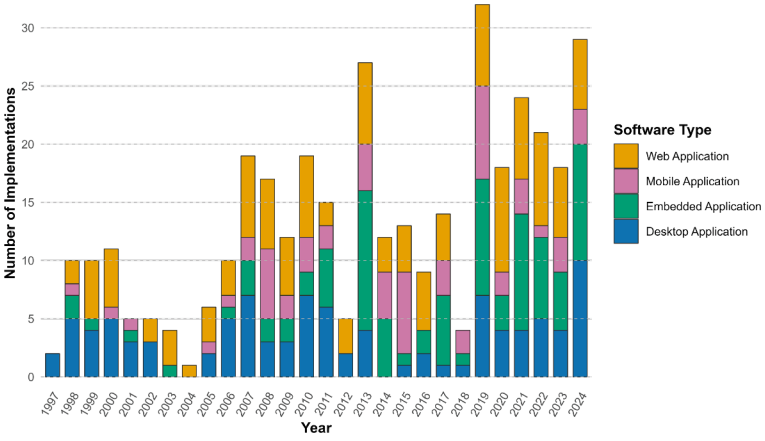


Fig. 2. Trend of software types in which IUIs are implemented, between 1997 and 2024.

From the review of articles, several combinations of software development process phases have been identified. Figure 3 summarizes the number of articles that refer to different combinations of software process phases in which IUIs are implemented.

The triad "Modeling, Construction, Deployment" (27.22%) and the tetrad "Communication, Modeling, Construction, Deployment" (24.36%) are the most frequent *co-occurrence patterns of lifecycle stages* across the articles. *Per-stage frequencies* (Communication, Modeling, Construction, Deployment) are reported separately in Table 5. These results indicate sustained attention to end-to-end IUI implementation, from conceptualization through deployment, with communication as a critical phase.

On the other hand, the combination "Modeling, Construction," mentioned in 19.77% of the articles, confirms the fundamental relationship between these two phases. In contrast, combinations that incorporate the "Communication" phase are less frequent, indicating that this phase is considered of lower relevance in some articles.

In summary, communication with users and stakeholders is the phase of the software development process in which the implementation of IUIs is addressed the least. This constitutes a critical gap, as a deep understanding of user requirements is essential for designing effective IUIs. In contrast, the modeling and construction phases were the most addressed, motivating the need to adopt model-driven approaches to integrate communication in the early stages of the IUI development process.

3.4 Techniques Used for the Implementation of IUIs

Due to the diversity of techniques analyzed, categories were established considering synonyms, terminological ambiguities, and the context of each technique. As

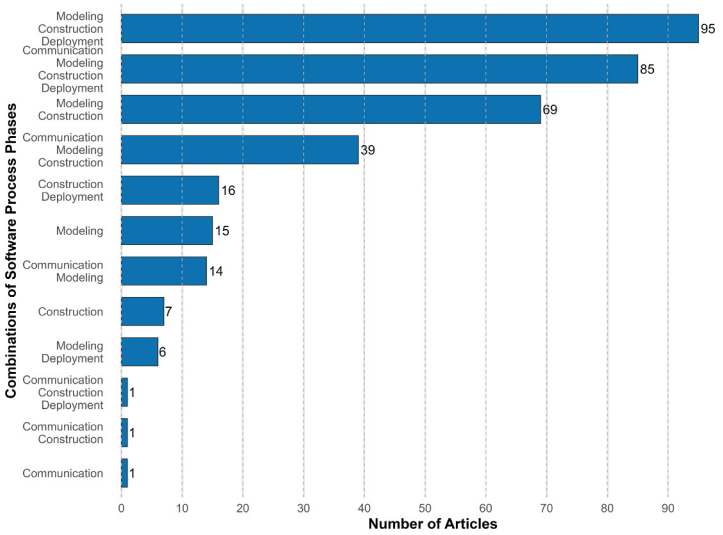


Fig. 3. Combinations of software process phases

a result of the analysis of Q4, groupings were defined to organize the identified techniques. Figure 4a shows the number of mentions for each technique category, highlighting Machine Learning and AI (27.6%) and Data and Signal Processing (27.1%) as the most used, followed by Algorithms and Models (23.1%) and Modeling and Ontologies (22%), underscoring their relevance in current research. The

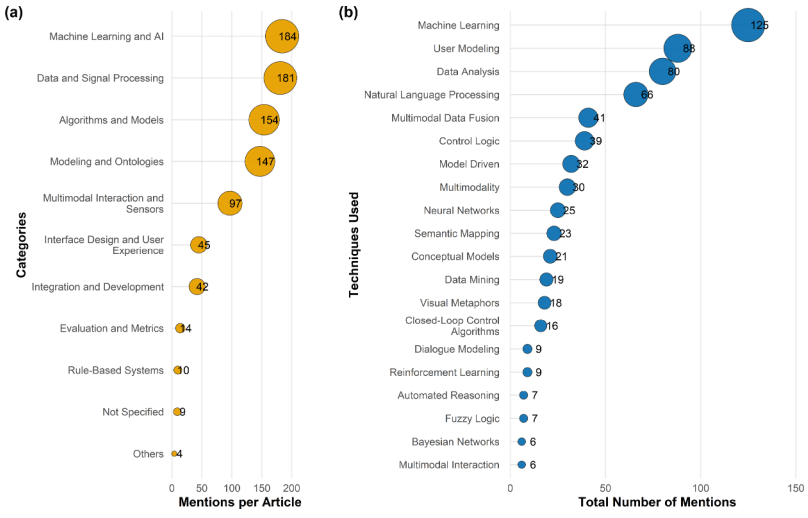


Fig. 4. Techniques for the implementation of IUIs

integration of techniques related to Multimodal Interaction and Sensors (14.5%) has a moderate presence, while categories such as Interface Design and User Experience (6.7%) and Integration and Development (6.3%) are less frequent.

In Fig. 4b, the most frequently used techniques for implementing IUIs are presented, considering only those with more than five mentions. Machine Learning is the most prevalent at 14.09%, followed by User Modeling at 9.92%, Data Analysis at 9.02%, and Natural Language Processing at 7.44%. These results reflect the importance of machine learning, user-based personalization, and data processing in the development of interactive interfaces.

Figure 5 shows the relationship between the most frequently used techniques and the types of software. Machine Learning (ML) has the highest presence in Web and Mobile Applications, representing 9.60% of total mentions, highlighting its importance in automating and personalizing interfaces. User Modeling (7.20%) and Natural Language Processing (6.75%) also stand out in these applications, reflecting their role in interface adaptation and user behavior analysis. In contrast, Multimodality (2.40%) has a greater presence in Embedded and Desktop Applications, indicating its relevance in interaction through multiple channels. Finally, Semantic Mapping (1.65%) shows lower representation in Web and Mobile Applications, although it remains a key technique in data structuring and comprehension.

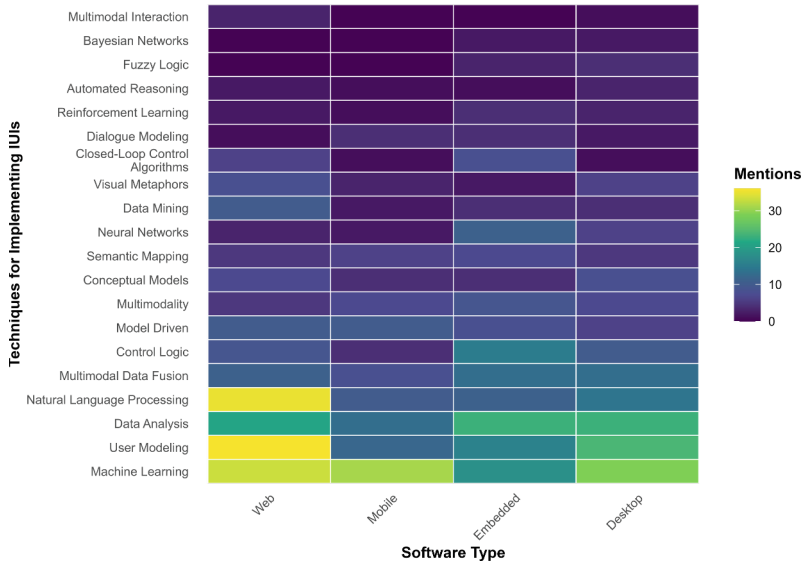


Fig. 5. Relationship between IUI implementation techniques and types of software

In summary, the development of IUIs requires a multidimensional approach. Although ML is the most widely used technique in IUI implementation, its application is not isolated; it is combined with data and signal processing and mul-

timodal interaction to enhance interface adaptability and personalization. On the other hand, techniques such as user modeling and natural language processing complement artificial intelligence. The heatmap analysis reveals that these techniques are predominant in web and mobile applications, while multimodal interaction is more relevant in embedded environments.

3.5 Quality Characteristics Associated with the Incorporation of IUIs

The analysis of the results for Q5 shows that Interaction Capability(22.09%) is the most frequently mentioned characteristic. Functional Suitability (21.35%) and Performance Efficiency (20.68%) also have a strong presence. Flexibility (14.60%) is relevant for software adaptability. Maintainability (9.93%) and Reliability (8.45%) are less frequent but remain key. On the other hand, Security (2.74%) is the least considered in the analyzed articles.

In summary, interoperability and functional suitability are the most addressed quality attributes, aligning with the central objective of IUIs: *enhancing the user experience*. However, security and safety have been scarcely considered, which represents a risk in critical environments such as IoT and healthcare, where data reliability and privacy are essential.

4 Conclusion and Future Work

This preliminary study examines the implementation of IUIs in software development and their impact on quality, from 2019 to 2024. A total of 366 articles were analyzed, and the findings indicate that IUI implementation is mainly based on Machine Learning, complemented by data processing, user modeling, and multimodal interaction.

In terms of quality, IUIs have been shown to enhance interaction and functional suitability. However, critical aspects such as security and reliability have been scarcely addressed, posing challenges in sensitive environments such as IoT and healthcare. A considerable number of the analyzed articles do not involve the development of the IUI within the software development process, which hinders the systematization of IUI development.

Furthermore, a growing trend toward implementation in web applications and mobile devices was observed, suggesting a preference for more accessible and scalable platforms. It is important to note that in this field there is terminological ambiguity between adaptive and intelligent interfaces, highlighting the need to consolidate standardized definitions in the area. Likewise, the limited attention to the communication phase and the lack of evaluation metrics restrict the industrial adoption of IUIs.

As future work, the analysis will be extended to examine the evaluation techniques used for IUIs and how these contribute to quality attributes. These advances will foster the development of more robust, accessible IUIs that are aligned with the demands of human interaction in increasingly complex environments.

5 Availability of Artifacts

Study list (titles/DOIs), selection/extraction matrix, and replication materials: Zenodo https://doi.org/10.5281/zenodo.17283843.

Acknowledgments. This research work was supported by the Escuela Politécnica Nacional (Ecuador) under PIIF-24-02 and PII-DICC-2025-01 research projects.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abrahão, S., Insfran, E., Sluÿters, A., Vanderdonckt, J.: Model-based intelligent user interface adaptation: challenges and future directions. Softw. Syst. Model. **20**(5), 1335–1349 (2021). https://doi.org/10.1007/s10270-021-00909-7
2. Ali, M., Khan, S.U.R., Mashkoor, A., Taskeen, A.: A conceptual framework for context-driven self-adaptive intelligent user interface based on android. Cogn. Technol. Work **26**(1), 83–106 (2024). https://doi.org/10.1007/s10111-023-00749-z
3. Association for Computing Machinery: ACM digital library. https://dl.acm.org (2025)
4. Brdnik, S., Heričko, T., Šumak, B.: Intelligent user interfaces and their evaluation: a systematic mapping study. Sensors **22**(15) (2022). https://doi.org/10.3390/s22155830, https://www.mdpi.com/1424-8220/22/15/5830
5. Brdnik, S., Šumak, B.: Adaptive and intelligent user interfaces and how they adapt to users: a systematic mapping study. In: Proceedings of the Faculty of Organization and Informatics Varazdin, pp. 327–334. Faculty of Organization and Informatics Varazdin (2022). http://archive.ceciis.foi.hr/app/index.php/ceciis/archive
6. Budgen, D., Turner, M., Brereton, P., Kitchenham, B.A.: Using mapping studies in software engineering. In: Annual Workshop of the Psychology of Programming Interest Group (2008). https://api.semanticscholar.org/CorpusID:1555016
7. Cedillo, P., Sanchez-Zhunio, C., Bermeo, A.: A systematic mapping study for intelligent user interfaces. In: 2017 International Conference on Information Systems and Computer Science (INCISCOS) (2017). https://doi.org/10.1109/INCISCOS.2017.34
8. Clarivate: Web of science. https://www.webofscience.com (2025)
9. Ehlert, P.: Intelligent user interfaces: introduction and survey. Research Report DKS03-01 / ICE 01, Delft University of Technology, The Netherlands (2003)
10. Elsevier: Scopus. https://www.scopus.com (2025)
11. Gonçalves, T.G., Kolski, C., de Oliveira, K.M., Travassos, G.H., Strugeon, E.G.L.: A systematic literature review on intelligent user interfaces: preliminary results. In: Adjunct Proceedings of the 31st Conference on l'Interaction Homme-Machine. IHM '19 Adjunct, Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3366551.3370344
12. Gonçalves, T.G., da Rocha, A.R.C.: Development process for intelligent user interfaces: an initial approach. In: Proceedings of the XVIII Brazilian Symposium on Software Quality, pp. 210–215. SBQS '19, Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3364641.3364665

13. Hartmann, M.: Challenges in developing user-adaptive intelligent user interfaces. In: Proceedings of the 17th Workshop on Adaptivity and User Modeling in Interactive Systems. Köck, Mirjam (2009). http://abis.l3s.uni-hannover.de/images/proceedings/abis2009/abis2009_hartmann.pdf, https://dl.gi.de/handle/20.500.12116/5024
14. Höök, K.: Steps to take before intelligent user interfaces become real. Interact. Comput. **12**(4), 409–426 (2000). https://doi.org/10.1016/S0953-5438(99)00006-5, https://www.sciencedirect.com/science/article/pii/S0953543899000065
15. IEEE: Iso/iec/ieee international standard - systems and software engineering–vocabulary. ISO/IEC/IEEE 24765:2017(E), pp. 1–541 (2017). https://doi.org/10.1109/IEEESTD.2017.8016712
16. IEEE: IEEE xplore digital library. https://ieeexplore.ieee.org (2025)
17. International Organization for Standardization (ISO): ISO/IEC 25010:2023 - Systems and Software Quality Requirements and Evaluation (SQuaRE). International Standard (2023). https://www.iso.org/obp/ui/es/#iso:std:iso-iec:25010:ed-2:v1:en, Último acceso: febrero 2024
18. Jaquero, V.L., Montero, F., Molina, J., GonzÂťlez, P.: Intelligent user interfaces: past, present and future, pp. 1–12. Springer, London (2009). https://doi.org/10.1007/978-1-84800-136-7_18
19. Kitchenham, B.: Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report Version 2.3, Keele University, UK (2007)
20. Klotins, E., Unterkalmsteiner, M., Gorschek, T.: Software engineering knowledge areas in startup companies: a mapping study. In: Fernandes, J.M., Machado, R.J., Wnuk, K. (eds.) Software Business, pp. 245–257. Springer International Publishing, Cham (2015)
21. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). https://doi.org/10.2307/2529310
22. Maybury, M.: Intelligent user interfaces: an introduction. In: Proceedings of the 4th International Conference on Intelligent User Interfaces, pp. 3–4. IUI '99, Association for Computing Machinery, New York (1998). https://doi.org/10.1145/291080.291081
23. Mezhoudi, N., Khaddam, I., Vanderdonckt, J.: Toward usable intelligent user interface. In: Kurosu, M. (ed.) Human-Computer Interaction: Interaction Technologies, pp. 459–471. Springer International Publishing, Cham (2015)
24. Moola, S., et al.: Conducting systematic reviews of association (etiology): the joanna briggs institute's approach. Int. J. Evidence-based Healthc. **13** (08 2015). https://doi.org/10.1097/XEB.0000000000000064
25. Motta, R.C., de Oliveira, K.M., Travassos, G.H.: On challenges in engineering Iot software systems. In: Proceedings of the XXXII Brazilian Symposium on Software Engineering, pp. 42–51. SBES '18, Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3266237.3266263
26. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, vol. 17 (2008)
27. Pressman, R.S., Maxim, B.R.: Ingeniería del software: un enfoque práctico. McGraw-Hill Interamericana, México, novena edición edn. (2021), traducción del original: Software Engineering: A Practitioner's Approach
28. Sanchez, C., Cedillo, P., Bermeo, A.: A systematic mapping study for intelligent user interfaces - IUI. In: 2017 International Conference on Information Systems and Computer Science (INCISCOS), pp. 361–368 (2017). https://doi.org/10.1109/INCISCOS.2017.34

29. Springer Nature: Springerlink. https://link.springer.com (2025)
30. Völkel, S.T., Schneegass, C., Eiband, M., Buschek, D.: What is "intelligent" in intelligent user interfaces? a meta-analysis of 25 years of IUI. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 477–487. IUI '20, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3377325.3377500

Check for updates

Semantic Search for Financial Products Using Embeddings and Pgvector

Esteven Calcina-Puma(✉) and Antonio Arroyo-Paz(✉)

Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru
{ecalcinap,aarroyop}@unsa.edu.pe

Abstract. This article presents a semantic search system for financial products using vector embeddings and PostgreSQL with the pgvector extension. The all-MiniLM-L6-v2 model transforms attributes such as account type and currency into vector representations, facilitating searches based on semantic similarity. An automated pipeline was implemented that includes web scraping of data from the website of the Superintendency of Banking, Insurance, and Pension Fund Administrators (SBS) of Peru, followed by processing with Django and a frontend in React. The results show excellent performance in single-criterion queries, obtaining a Recall@10 close to 1.0, which indicates that the relevant results are among the top ten. In queries with combined criteria, such as entity plus location, Recall@10 reaches 0.5. It is concluded that the automation of data collection and processing, together with the generation of semantic embeddings and their efficient storage in PostgreSQL using pgvector, allowed the development of a system capable of interpreting queries in natural language and surpassing traditional methods, demonstrating high performance in tests with real data.

Keywords: Semantic Search · Vector Embeddings · Pgvector · NLP Applied to Finance · Banking Product Recommendation · SentenceTransformer

1 Introduction

The digital transformation of the financial sector has increased the complexity and diversity of banking products available to users, which represents a significant challenge in terms of discovery and personalized recommendation [1]. In a technical report on the evaluation of data management and analysis systems, the Superintendency of Banking and Securities (SBS) [2] points out that one of the requirements (EU07) for its search solution is the ability to perform searches using natural language, which contrasts with traditional systems based on structured filters and lexical matching [3]. On the other hand, there are web alternatives that allow the search and comparison of financial products offered by various entities in Peru. One of them is the platform https://comparabien.com.pe [4], which allows the comparison of different banking products through

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 358–373, 2026.
https://doi.org/10.1007/978-3-032-20322-9_25

simple forms with basic fields. Likewise, the SBS offers an online tool [5] to compare savings product rates, which also uses a single, simple form structure.

In this context, semantic search [6], based on representing meaning using high-dimensional vectors, has established itself as an effective alternative to traditional approaches [7]. Recent techniques in Natural Language Processing (NLP), such as Sentence Embedding models [8], allow complex texts to be converted into vector representations that capture semantic relationships beyond lexical matches [7]. Models such as all-MiniLM-L6-v2 [9], developed by Microsoft Research, offer an optimal balance between accuracy and computational efficiency, achieving remarkable scores in benchmarks such as STS (Semantic Text Similarity) [10].

Furthermore, the development of vector extensions for relational databases, such as pgvector for PostgreSQL [11], has made it possible to incorporate cosine similarity searches directly into storage systems, without requiring external search engines such as Elasticsearch or FAISS [12]. This simplifies the architecture and enables scalability in systems that require semantic retrieval.

Previous studies have shown that retrieval models based on dense embeddings can outperform traditional systems such as TF-IDF (Term Frequency – Inverse Document Frequency) and BM25 (Best Match 25) by up to 19% in information retrieval tasks [7]. While TF-IDF weights the relevance of a word in a document according to its frequency in the text and its rarity in the corpus, BM25 improves this approach by introducing term frequency saturation and length normalization. These techniques work well for lexical matches but lack semantic understanding, limiting their effectiveness compared to embedding-based methods, particularly in specialized domains such as e-commerce, healthcare, and finance.

Given the above, this paper presents the development of a semantic search system capable of retrieving financial savings products from the SBS portal using natural language queries, embeddings generated with the all-MiniLM-L6-v2 model, and PostgreSQL with pgvector.

In this context, a *savings financial product* is defined as an account or deposit offered by institutions in the Peruvian banking system, characterized by structured attributes such as: the `entity` offering it, the `location` where it is available, the associated annual percentage rate (APR), the account type (e.g., savings or term deposit), the currency in which it is offered (soles or dollars), and the specific conditions or requirements.

2 Related Works

Nigam et al. [13] address the problem of semantic search in e-commerce, where traditional lexical match-based engines fail to capture synonyms, hypernyms, or spelling errors. They propose a deep learning model with shared embeddings, a "3-part hinge" loss function to differentiate between purchased, viewed, and random products, and hybrid tokenization (unigrams, n-grams, and character trigrams). Trained on 54 billion Amazon query–product pairs, their model achieves

a Recall@100 of 79.4% and MAP of 74.5%, surpassing methods such as DSSM and MatchPyramid by 14.5%. Online tests showed significant increases in conversion rates and revenue. Their approach stands out for combining simplicity (average pooling) and scalability (distributed training on 8 GPUs through cosine similarity decomposition), offering a framework applicable to domains such as financial search, although with adaptations for structured data and specific colloquial queries.

J. Malmberg [14] addresses the problem of semantic similarity search through sentence embeddings, comparing traditional methods (TF-IDF) with neural approaches (BERT and Sentence-BERT). The study evaluates performance on modified versions of the SQuAD 1.0 datasets (for question–answer matching) and STS-B (for sentence pair similarity), using mean-pooling strategies and the [CLS] token to generate BERT-based embeddings. Results show that TF-IDF outperforms neural models in most scenarios (77.2% accuracy compared to SBERT's 72.9%), especially when using full-dimension vectors, although SBERT outperforms standard BERT. Key findings highlight TF-IDF's robustness in contexts with strong lexical similarity, while SBERT's siamese architecture improves semantic alignment. The work notes limitations due to the lack of domain-specific datasets and recommends evaluating modern architectures such as ELECTRA.

Reimers and Gurevych [15] address the limitations of BERT and RoBERTa in sentence semantic similarity tasks, noting that although these models achieve high performance in regression tasks such as STS (Semantic Textual Similarity), they require both sentences to be processed jointly, implying a high computational cost (around 50 million inferences to compare 10,000 sentences). To solve this problem, the authors propose Sentence-BERT (SBERT), a modification of BERT that employs siamese and triplet architectures to generate comparable sentence embeddings via cosine similarity. This structure allows for a drastic reduction in semantic search time (from 65 h to 5 s in the same scenario) while maintaining BERT's accuracy. Furthermore, SBERT and its SRoBERTa variant outperform other sentence embedding methods in standard STS and transfer learning tasks, highlighting their effectiveness in both efficiency and semantic quality.

Gupta et al. [16] address the challenge of improving document retrieval and classification in information systems through a comparative analysis of sentence embedding techniques. They evaluate six pre-trained models: SentenceBERT, Universal Sentence Encoder (USE), InferSent, ELMo, XLNet, and Doc2Vec, using four standard datasets (CACM, CISI, ADI, and Medline). Results show that USE and SentenceBERT outperform the other models in key metrics such as MAP (Mean Average Precision), recall, F-measure, and NDCG, highlighting their ability to capture semantic and contextual similarities. In particular, USE achieves superior performance in three of the four datasets, while SentenceBERT stands out in Medline. The study demonstrates the effectiveness of these models for information retrieval tasks and suggests their potential in applications such as query reformulation and hybrid classification systems.

Coelho et al. [17] tackle the challenge of improving semantic search in mobile applications by using word embedding models. They compare word-level embeddings (GloVe), character-level embeddings (FastText), and contextual embeddings (RoBERTa), trained and fine-tuned with a dataset of 500,000 applications. Evaluation is performed on a public dataset with 43,000 apps and 56 non-exact queries, using the NDCG metric. Results show that FastText trained from scratch (FT2) and fine-tuned RoBERTa (RoBERTapp) outperform pre-trained models and traditional methods like BM25, with NDCG@5 of 0.589 and 0.587 respectively. They particularly excel in semantic queries like "social network," where RoBERTapp retrieves relevant apps (Facebook, Instagram) that lexical models do not prioritize. The study demonstrates that contextual and character embeddings significantly improve accuracy in non-exact searches, opening possibilities for smarter recommendation engines.

Asmala [18] addresses the challenge of improving the search of medical articles using Sentence Transformer models. The study focuses on optimizing scientific information retrieval by converting medical texts into semantic vectors, enabling more accurate searches based on meaning similarity rather than traditional keywords. To this end, different transformer architectures, such as BERT and biomedicine-specialized models (e.g., BioBERT), are evaluated, comparing their performance in document retrieval tasks. Standard medical datasets are used to validate results, measuring metrics such as precision and recall. Findings show that fine-tuned Sentence Transformers in medical domains significantly outperform classic search methods, highlighting their potential to streamline scientific research. The proposed solution integrates these models into semantic search pipelines, offering an efficient alternative to traditional lexical index-based systems.

Consuegra Pérez [19] addresses the development of a generative AI-based chatbot to improve productivity in fashion stores using the RAG (Retrieval-Augmented Generation) technique, combining a language model (GPT-3.5 Turbo) with a vector database of internal documents. The study evaluates the system's effectiveness in nine pilot stores of the Tendam group, measuring the precision of generated answers to queries about processes, manuals, and regulations, as well as the impact on reducing information search times. Results show that the chatbot achieves 85% satisfaction in relevant answers when tags (TAGs) are used to avoid ambiguities, although limitations such as uncontrolled responses to undocumented topics are identified. The conclusions highlight the feasibility of the RAG approach for multinational retail environments, with an estimated annual ROI of €609,465, and propose future research lines such as integration with structured data and image-based search. The work underscores the importance of document quality and embedding parameterization (text-embedding-3-small) to optimize semantic retrieval.

Acosta Bernal [20] proposes a generative AI platform to combat child malnutrition in Colombia using the RAG (Retrieval-Augmented Generation) technique, integrating language models (GPT-4) with a PostgreSQL/pgvector vector database storing BERT-generated embeddings. The study addresses the problem

of fragmented nutritional information and the lack of tools for its analysis, focusing on classifying academic articles, generating personalized recommendations, and facilitating decision-making. Results show that analytical queries achieve 97% accuracy compared to 47% for direct questions, highlighting pgvector's effectiveness in semantic retrieval of the 88,900 processed documents. The conclusions emphasize the solution's potential for public policy, with an estimated saving of 200,000 annual hours in information search, and identify key challenges such as expansion to 600,000 documents and improving handling of ambiguous queries. The work emphasizes the importance of multilingual embeddings and HNSW optimization for complex queries in resource-constrained environments.

3 System Implementation

As shown in Fig. 1, the system automates data collection from the SBS website through web scraping, processes this information, and generates semantic embeddings of financial products using the all-MiniLM-L6-v2 model. These embeddings are stored in PostgreSQL using the pgvector extension to enable efficient cosine similarity searches. A semantic search is implemented that connects natural language queries with relevant products, a frontend is developed for user interaction, and the system is validated with real data from Peruvian financial institutions.

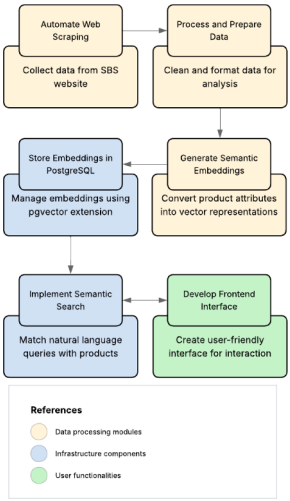


Fig. 1. Components and processes available within the Financial Data Analysis System (FDAS)

3.1 Architecture

Figure 2 shows the architecture of the system that was implemented.

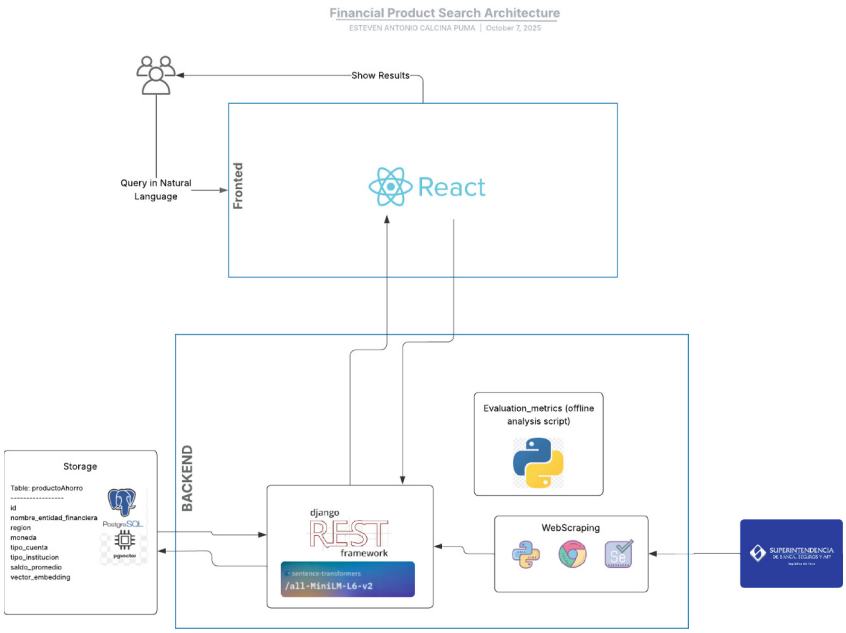


Fig. 2. System architecture

Frontend—React 19.1.0: The user interface was developed in React [21] in a local environment, on a computer running the UBUNTU operating system, allowing users to make queries in natural language through a simple form. Requests are sent to the backend via HTTP (POST) requests, and the results are dynamically rendered as a list of recommended products, sorted by similarity.

Backend—The data collection process was successfully automated using web scraping techniques, including the incorporation of a script developed with Selenium (version 4.34.2) [22] that allows dynamic interaction with web pages to obtain the necessary information. This automation facilitated the construction and constant updating of a database with financial products from various Peruvian institutions, eliminating the need for manual intervention and improving the efficiency and accuracy of data capture.

Subsequently, the information obtained was processed and treated to ensure its quality and uniformity. Cleaning, standardization, and structuring tasks were performed on key attributes such as account type, currency, region, and minimum balance, preparing the data for semantic analysis.

The logic is concentrated in **Django 4.2.23** [23] + **SentenceTransformers 3.2.1**. The system resides in an API built with *Django REST Framework* and

was tested in a local environment on a computer running the *Ubuntu* operating system. The version of *Python* used is 3.8.10. This layer exposes two main functionalities:

Financial Product Registration: This process is carried out by running a script that inserts all the data obtained through *web scraping* from the SBS portal. Each product has attributes such as *región*, *moneda*, *tipo de cuenta*, *tipo de institución*, and *saldo promedio*, which are stored in a table called `productosAhorro`. During registration, the product's textual attributes are concatenated and transformed into a vector using the *all-MiniLM-L6-v2* model from *SentenceTransformers*. This 384-dimensional vector is stored in the `vector_embedding` column within the same table, using the *VectorField* data type provided by *pgvector* in PostgreSQL.

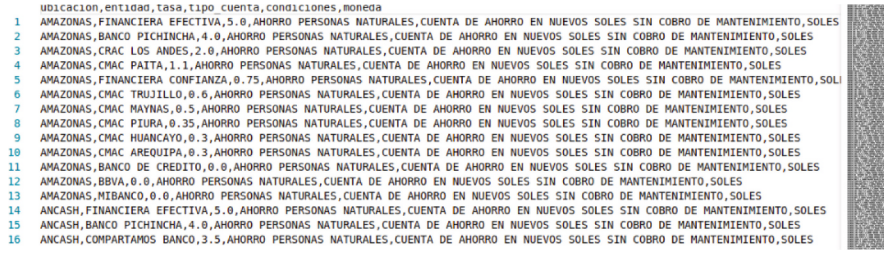
Semantic Query: upon receiving a textual query from the user, the backend generates a corresponding embedding vector and performs a search in the database using cosine similarity comparison. The all-MiniLM-L6-v2 model is loaded only once when the server starts, optimizing encoding performance.

Database—PostgreSQL 17.5 + Pgvector 0.3.6: Product storage and semantic vectors are managed in PostgreSQL in a local environment, on a machine running the UBUNTU operating system, extended with the pgvector library. This allows vector-type fields to be defined and approximate searches to be performed using specialized operators such as <=>. HNSW indexes are implemented to speed up approximate nearest neighbor searches [24].

3.2 Development

The proposed system was developed as a modular architecture based on open source technologies, integrating natural language processing, vector storage, and an API-type web interface. The solution allows users to query financial products using natural language, returning relevant results based on the semantic similarity between the query and the product descriptions. First, a web scraping process was performed on the official SBS website https://www.sbs.gob.pe/app/retasas/paginas/retasasInicio.aspx?p=D using Selenium and Python scripts, which allowed for dynamic and reliable data extraction. Subsequently, the collected data was processed using specific scripts to standardize records and consolidate the information into a structured format, reaching a total of 1,528 records, as shown in Fig. 3.

Finally, this clean and structured information was used to populate the database, generating semantic embeddings with the all-MiniLM-L6-v2 model that represented each financial product in vector form, thus enabling efficient and relevant searches based on semantic similarity.



```
ubicacion,entidad,tasa,tipo_cuenta,condiciones,moneda
AMAZONAS,FINANCIERA EFECTIVA,5.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,BANCO PICHINCHA,4.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CRAC LOS ANDES,2.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CMAC PAITA,1.1,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,FINANCIERA CONFIANZA,0.75,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOL
AMAZONAS,CMAC TRUJILLO,0.6,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CMAC MAYNAS,0.5,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CMAC PIURA,0.35,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CMAC HUANCAYO,0.3,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,CMAC AREQUIPA,0.3,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,BANCO DE CREDITO,0.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,BBVA,0.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
AMAZONAS,MIBANCO,0.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
ANCASH,FINANCIERA EFECTIVA,5.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
ANCASH,BANCO PICHINCHA,4.0,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
ANCASH,COMPARTAMOS BANCO,3.5,AHORRO PERSONAS NATURALES,CUENTA DE AHORRO EN NUEVOS SOLES SIN COBRO DE MANTENIMIENTO,SOLES
```

Fig. 3. Data obtained from WebScraping

Embedding Generation: To transform structured information from financial products into vector representations, the pre-trained model all-MiniLM-L6-v2 from the SentenceTransformers repository was used. This model, optimized for semantic similarity tasks, converts multi-component texts into 384-dimensional vectors. Semantic encoding considers key attributes such as ubicacion, tipo_cuenta, and moneda, concatenated in a template-like text structure. An example of the input format is shown in Listing 1.1.

Listing 1.1. Input in JSON format

```
{
  "ubicacion": "ayacucho",
  "entidad": "BANCO PICHINCHA",
  "tasa": 4.0,
  "tipo_cuenta": "Ahorro tipo Natural",
  "condiciones": "sin mantenimiento",
  "moneda": "PEN"
}
```

The model is invoked from a custom view in Django, in the perform_create and perform_update methods of the ProductoAhorroSBSViewSet ViewSet. Each time a new product is registered or modified, its embedding is generated and stored directly in the vector_embedding field of the model. The database population was automated with a script, resulting in Fig. 4.

Vector Storage with Pgvector: The generated vectors are stored in a PostgreSQL database using the pgvector extension. This extension allows you to define vector-type fields and perform similarity searches using operators such as <=> (cosine distance). The products_productsavings table includes a vector index to speed up approximate nearest neighbor (ANN) searches, enabling scalability for larger data volumes.

Semantic Query API: A specific endpoint for vector searches (BusquedaVectorialSBSAPIView) was implemented, which receives a textual query from the user and transforms it into a vector using the `all-MiniLM-L6-v2` model. Subsequently, this vector is compared against the records stored in the database

```
(ambiente) stv@stv-Inspiron-15-3520:~/TopicosAvanzadoDb/proyecto/buscadorBack$ python llenarBaseDatos.py

REGISTRO DE PRODUCTOS FINANCIEROS

N°   Entidad                    Ubicación    Tasa   Tipo de Cuenta                Moneda   Estado
-------------------------------------------------------------------------------------------------
1    FINANCIERA EFECTIVA        AMAZONAS     5.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
2    BANCO PICHINCHA            AMAZONAS     4.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
3    CRAC LOS ANDES             AMAZONAS     2.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
4    CMAC PAITA                 AMAZONAS     1.1    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
5    FINANCIERA CONFIANZA       AMAZONAS     0.75   AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
6    CMAC TRUJILLO              AMAZONAS     0.6    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
7    CMAC MAYNAS                AMAZONAS     0.5    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
8    CMAC PIURA                 AMAZONAS     0.35   AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
9    CMAC HUANCAYO              AMAZONAS     0.3    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
10   CMAC AREQUIPA              AMAZONAS     0.3    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
11   BANCO DE CREDITO           AMAZONAS     0.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
12   BBVA                       AMAZONAS     0.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
13   MIBANCO                    AMAZONAS     0.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
14   FINANCIERA EFECTIVA        ANCASH       5.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
15   BANCO PICHINCHA            ANCASH       4.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
16   COMPARTAMOS BANCO          ANCASH       3.5    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
17   CMCP LIMA                  ANCASH       1.0    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
18   FINANCIERA CONFIANZA       ANCASH       0.75   AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
19   CMAC DEL SANTA             ANCASH       0.6    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
20   CMAC TRUJILLO              ANCASH       0.6    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
21   CMAC PIURA                 ANCASH       0.35   AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
22   CMAC HUANCAYO              ANCASH       0.3    AHORRO PERSONAS NATURALES     SOLES    ✅ Registrado
^CTraceback (most recent call last):
```

Fig. 4. Datase input.

through a raw SQL query (see Listing 1.3) that uses the `<=>` operator to compute cosine similarity. The query retrieves the most representative fields of each product: `ubicacion`, which indicates the region or city where the savings product is offered; `entidad`, corresponding to the financial institution providing the product; `tasa`, representing the annual effective interest rate associated with it; `tipo_cuenta`, which identifies the type of product (for example, savings account or term deposit); `condiciones`, which describes the specific requirements or benefits of the product; and `moneda`, which specifies whether the account is denominated in soles or dollars. Finally, the `embedding` field contains the vector representation generated from these attributes, which enables the computation of semantic similarity between products. The results are sorted by vector proximity, and the five most relevant products are returned along with their main characteristics.

Vector storage is managed with the `pgvector` extension, which allows you to define a specialized field for embeddings. The schema of the `products` table is shown in Listing 1.2, where you can see that each record includes structured fields (region, currency, account type, etc.) and a `vector_embedding` field of type `vector(384)` that stores the semantic representation of the product.

Listing 1.2. Schema of the product table for vector storage in PostgreSQL with pgvector.

```
CREATE TABLE productoAhorro (
    id SERIAL PRIMARY KEY,
    region VARCHAR(20) DEFAULT 'lima',
    moneda VARCHAR(10) DEFAULT 'soles',
    tipo_cuenta VARCHAR(10) DEFAULT 'ahorro',
    tipo_institucion VARCHAR(15) DEFAULT 'banco',
    saldo_promedio NUMERIC(15,2),
    vector_embedding VECTOR(384)
);
```

Figure 5 shows an example of a semantic query performed in Postman. This endpoint is provided by the backend for later use in the frontend.

Listing 1.3. SQL query for products

```
SELECT id, ubicacion, entidad, tasa,
tipo_cuenta, condiciones, moneda, embedding
FROM
productos_productoahorrosbs
ORDER BY embedding <#>
%s::vector
LIMIT 10;
```

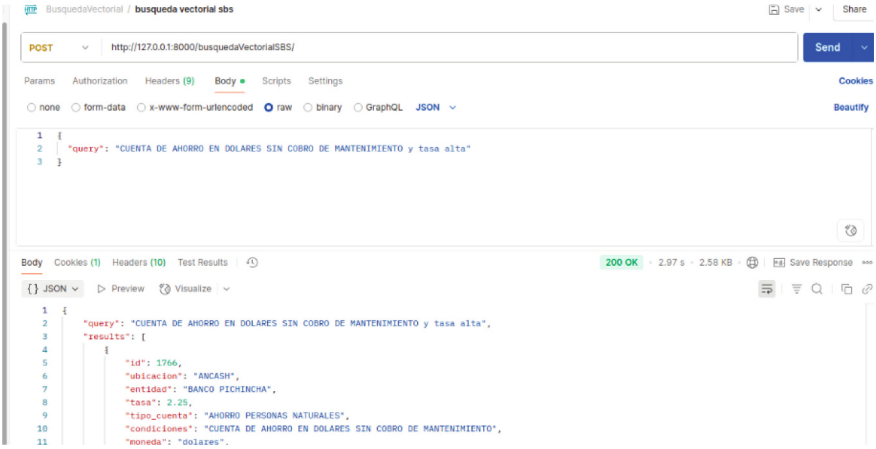


Fig. 5. Semantic query endpoint test.

Frontend: To facilitate end-user interaction with the semantic search system, a frontend web application was developed using React 18 and Vite as the build tool. The interface implements a responsive design based on Tailwind CSS and shadcn/ui components, offering a modern and accessible user experience. The main component includes a smart search field that allows the user to enter queries in natural language (e.g., "savings account in soles with no maintenance fees in Lima"), which are sent via HTTP requests to the vector search endpoint of the Django backend. The results are presented in information cards that show the interest rate, account type, bank, conditions, and location of each financial product found. Additionally, a loading status and error handling system was implemented to improve the user experience during queries, ensuring immediate visual feedback and clear communication of any connection problems with the API. Samples can be seen in Fig. 6.

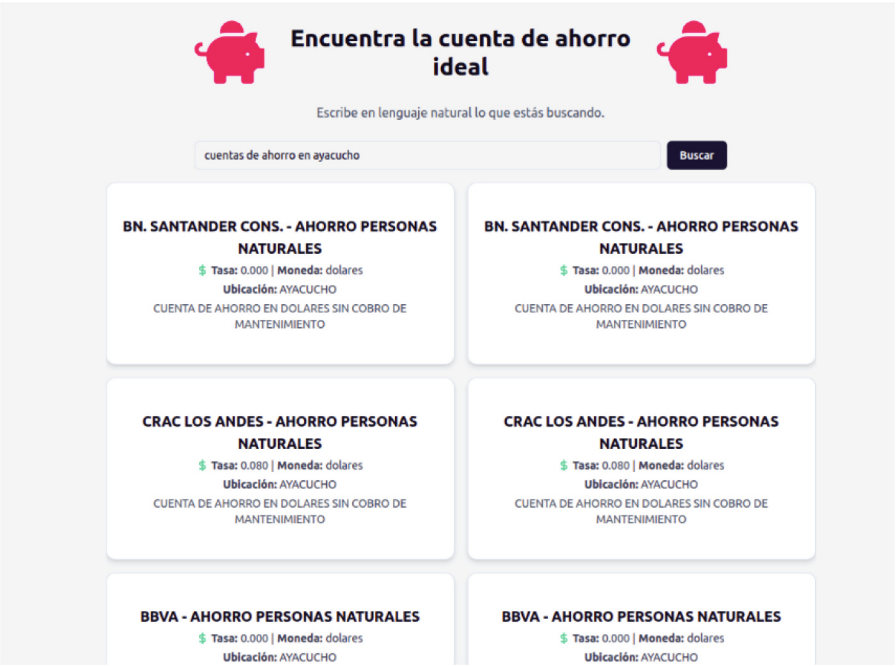


Fig. 6. Savings account search.

Integration and Deployment: The complete system was integrated into a Django REST Framework application, facilitating its use by frontend applications or conversational assistants. Consistency was maintained in text encoding for both insertion and querying to ensure semantic alignment. The solution can be deployed locally or on cloud servers, allowing for future integration with web or mobile interfaces.

4 Results

Two metrics were used to evaluate the semantic retrieval system: **Precision@k** and **Recall@k**, defined in Eqs. 1 and 2. However, the most relevant metric in this study is **Precision@k**, since queries are made by human users who generally expect accurate results within a small set of responses. In this work, $k = 10$ was established, considering that a list of ten results is adequate for manual review and to represent a realistic search scenario. In this context, G represents the set of relevant results, while the numerator expresses the intersection between the retrieved results (R_k) and the relevant ones G. A document is considered relevant when it meets a specific defined criterion. For example, in query Listing 1.4, we have the query **‘‘`cuentas de ahorro en ayacucho`’’**, all documents containing the location `Ayacucho` are considered relevant.

Listing 1.4. Entrada en formato JSON

```
{
    "query": "cuentas de ahorro en ayacucho",
    "criterio": "ubicacion",
    "relevantes": ["AYACUCHO"]
}
```

$$\text{Precision@k} = \frac{|r \in R_k \cap G|}{|R_k|} \tag{1}$$

where:

- R_k: Set of the top-k results returned by the system
- G: Set of all relevant results (ground truth)
- $R_k \cap G$: Intersection between retrieved and relevant results
- $|\cdot|$: Cardinality (number of elements in the set)

Recall@k, shown in Eq. 2, measures the proportion of relevant documents that were retrieved within the first k results returned by the system. Like Precision@k, the set R_k represents the financial savings products obtained by the search system in the top-k, while G corresponds to the set of relevant products previously defined as a reference or ground truth. The cardinality $|\cdot|$ indicates the number of elements in each set.

$$\text{Recall@k} = \frac{|r \in R_k \cap G|}{|G|} \tag{2}$$

where:

- R_k: Set of results returned in the top-k
- G: Set of expected relevant results (golden truth)
- $|\cdot|$: Cardinality (number of elements in the set)

The evaluation was performed using nine different queries on 1,527 records, as shown in Table 1, each corresponding to a different information need.

As shown in Table 1, the system's performance is evaluated using precision and recall metrics over the top 10 results (`Retrieved K = 10`) for each query (**Query**) and evaluation criterion (**Criterio**). The **Precision@10** column indicates the proportion of retrieved results that are truly relevant according to the set of relevant documents (**Relevant G**), while **Recall@10** reflects the fraction of relevant documents that were actually retrieved. **True Positives** represents the number of retrieved documents that match the relevant ones, **Retrieved K** is the number of documents considered in the evaluation (top-10), and **Relevant G** corresponds to the total number of relevant documents according to the reference set.

It is observed that the system achieves high performance in **ubicación** criteria, with **Precision@10** values of 0.8 and 1.0 for queries in Ayacucho and Arequipa, respectively, showing that most of the relevant documents appear

within the top 10 results. For the **entidad** criterion, some queries such as Cuentas de ahorro en la entidad BBVA" maintain high performance (**Precision@10 = 0.9**), whereas the query "Cuentas de ahorro del banco BBVA" obtains **Precision@10 = 0.0** and **Recall@10 = 0.0**, because the system confuses the mention of the institution with the generic word "banco", indicating an area for improvement in the interpretation of institution names.

Overall, these results show that the system is effective at retrieving relevant documents, especially for queries based on location and clearly identifiable entities, and highlight the need to refine the entity name recognition logic for ambiguous queries (Fig. 7).

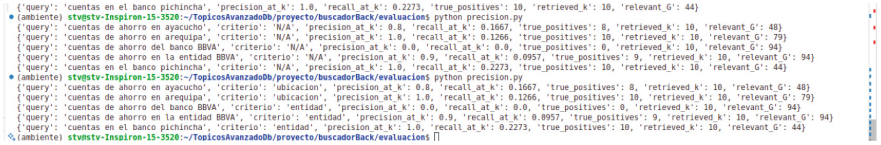


Fig. 7. Evaluation results.

Table 1. Extended Performance Metrics by Query and Evaluation Criterion

Query	Criterio	Precision@10	Recall@10	True Positives	Retrieved K	Relevant G
Cuentas de ahorro en ayacucho	ubicación	0.800	0.1667	8	10	48
Cuentas de ahorro en arequipa	ubicación	1.000	0.1266	10	10	79
Cuentas de ahorro del banco BBVA	entidad	0.000	0.0000	0	10	94
Cuentas de ahorro en la entidad BBVA	entidad	0.900	0.0957	9	10	94
Cuentas en el banco pichincha	entidad	1.000	0.2273	10	10	44
Cuentas de ahorro del Interbank en Arequipa	entidad + ubicación	0.500	0.5000	2	4	4
Cuentas de ahorro en Ayacucho del Banco de Crédito	entidad + ubicación	0.500	0.5000	2	4	4

5 Conclusion

The evaluation results presented in Table 1 show the overall effectiveness of the system in retrieving relevant financial products. Performance was measured using **Precision@10** and **Recall@10** over the top 10 retrieved results (`Retrieved K = 10`) for each query (**Query**) and evaluation criterion (**Criterio**). Precision

indicates the proportion of retrieved documents that are truly relevant according to the reference set (**Relevant G**), while recall reflects the fraction of all relevant documents that were actually retrieved.

The system achieves high performance in queries based on **ubicación**, with **Precision@10** values of 0.8 and 1.0 for Ayacucho and Arequipa, respectively, indicating that most of the relevant documents are found within the top 10 results. For the **entidad** criterion, queries such as "Cuentas de ahorro en la entidad BBVA" maintain high performance (**Precision@10 = 0.9**), whereas "Cuentas de ahorro del banco BBVA" obtains **Precision@10 = 0.0** and **Recall@10 = 0.0** because the system confuses the institution name with the generic word "banco". This highlights an opportunity for improvement in the identification and disambiguation of financial institution names.

These results indicate that the system is effective in retrieving relevant documents, especially for queries based on location and clearly identifiable entities, while also highlighting the need to refine the entity-recognition logic in ambiguous queries. As future work, it is recommended to improve entity disambiguation, incorporate more complex evaluation criteria, and extend the system to additional financial datasets.

It is observed that, in several queries, **Recall@10** is relatively low, indicating that only a fraction of all relevant documents is found within the top 10 results. This is mainly because the evaluation is limited to the top-10 and because some relevant documents do not appear in the highest-ranked positions, either due to ambiguity in entity names or partial matches in the search criteria. Despite this, greater importance is placed on **Precision@10**, since in practical applications it is preferable for users to first encounter highly relevant results, minimizing the appearance of non-relevant documents in the top positions of the list. This approach ensures a more satisfactory user experience, even if some relevant documents remain outside the evaluated top-k.

6 Discussion

This work is situated in the field of *information retrieval*, where datasets are not standardized and each database may present its own structures and criteria. This makes direct comparison with other approaches in the literature difficult and explains why metrics such as **recall** may be low: relevant documents are dispersed, and not all of them appear within the evaluated top-k, whereas **Precision@10** more clearly reflects the system's immediate effectiveness in delivering relevant results in the top positions, which is critical for financial applications.

Comparing with previous studies, Nigam et al. [13] achieve *Recall@100* of 79.4% and MAP of 74.5% in e-commerce, outperforming methods based on lexical matching. Although the datasets are different and there is no shared baseline—and moreover, that work places greater emphasis on recall, while in this work greater importance is given to precision—both approaches highlight the relevance of semantic embeddings and deep learning for capturing relationships between queries and products. Our system shows that, even with a proprietary and relatively small financial dataset, it is possible to retrieve highly

precise results within the first 10 documents, aligning with the idea of prioritizing immediate relevance.

Malmberg [14] and Reimers and Gurevych [15] show that models such as SBERT improve the efficiency and accuracy of semantic search through comparable embeddings, although methods based on the scientific concept TF-IDF can outperform neural models in contexts with high lexical overlap. This is reflected in our system: queries with clear lexical matches (for example, "Cuentas de ahorro en la entidad BBVA") maintain high **Precision@10**, while ambiguous queries rely more heavily on semantic capability, highlighting the need to combine lexico-semantic signals.

Gupta et al. [16] and Coelho et al. [17] highlight that pretrained embedding models (such as USE, SentenceBERT, and RoBERTa) outperform traditional methods in non-exact retrieval tasks. In our case, the retrieval of relevant documents depends on the quality of the embeddings and the correct interpretation of the query; semantic confusion reduces recall, as observed in "Cuentas de ahorro del banco BBVA".

Finally, RAG-based approaches combined with vector databases, such as those proposed by Consuegra Pérez [19] and Acosta Bernal [20], highlight the importance of embedding parameterization, document base structuring, and the optimization of vector indexes. Our system does not yet integrate entity disambiguation or hybrid query optimizations, which partially explains the low recall in ambiguous cases and suggests future improvements to maximize the retrieval of relevant documents without sacrificing the high **Precision@10** achieved.

In summary, although the financial datasets are proprietary and not directly comparable to benchmarks in the literature, the results show that our system achieves competitive performance—particularly in terms of precision—on simple and moderately complex queries, emphasizing the need to refine semantic interpretation [25].

References

1. América Retail, La transformaci ó n digital en la banca, personalizaci ó n de experiencias, https://america-retail.com/secciones/innovacion/la-transformacion-digital-en-la-banca-personalizacion-de-experiencias/, En línea (2024)
2. Superintendencia de Banca, Seguros y AFP, Home - portal sbs (2025). https://www.sbs.gob.pe/usuarios/. Accedido el 2 de agosto de 2025
3. J. K. Uechi, Informe no 00071-2021-gti-pri (2021). https://www.sbs.gob.pe/Portals/0/jer/INFTECSAO2021/INFORME%20N%2000071%20-%202021%20-%20GTI-PRI.pdf, Gerencia de Tecnologías de Información, Superintendencia de Banca, Seguros y AFP, Lima, Perú
4. Comparabien, Comparador de productos financieros en per ú (2025). https://comparabien.com.pe/. En línea. accedido el 7 de julio de 2025
5. Superintendencia de Banca, Seguros y AFP, Compara rendimiento de dep ó sitos - portal del usuario (2025). https://www.sbs.gob.pe/usuarios/informate-y-compara/compara-productos-financieros/compara-rendimiento-de-depositos. En línea. accedido el 7 de julio de 2025

6. P. Cudre-Mauroux, Semantic search (2025). https://link.springer.com/rwe/10.1007/978-3-319-63962-8_231-1. Accedido el 2 de agosto de 2025
7. Facebook AI Research, Dense passage retrieval for open-domain question answering (2025). https://arxiv.org/abs/2004.04906. Accedido el 7 de julio de 2025
8. Pilehvar, M.T., Camacho-Collados, J.: Embeddings in natural language processing: theory and advances in vector representations of meaning (synthesis lectures on human language technologies), vol. 47, pp. xvii+157. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-02177-0
9. Microsoft, All-minilm-l6-v2 - sentence transformers. https://www.sbert.net. En línea
10. Yin, C., Zhang, Z.: A study of sentence similarity based on the all-minilm-l6-v2 model with 'same semantics, different structure' after fine tuning. In: Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024), pp. 677–684. Atlantis Press (2024)
11. PostgreSQL, Pgvector extension. https://github.com/pgvector/pgvector. En línea
12. Knowi, Elasticsearch: What it is, how it works, and it's usage. https://www.knowi.com/blog/what-is-elastic-search/. En línea, accedido el 7 de julio de 2025
13. Nigam, P., Song, Y., Mohan, V., Lakshman, V., Ding, W., Shingavi, A., et al.: Semantic product search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), Anchorage, AK, USA, pp. 2876–2885 (2019)
14. Malmberg, J.: Evaluating semantic similarity using sentence embeddings. M.S. thesis, KTH Royal Institute of Technology, Estocolmo, Suecia (2021)
15. Reimers, N., Gurevych, I.: Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982–3992 (2019)
16. Gupta, V., Dixit, A., Sethi, S.: A comparative analysis of sentence embedding techniques for document ranking. J. Inf. Retr. **25**(4), 2155–2184 (2022)
17. Coelho, J., Neto, A., Tavares, M., Coutinho, C., Ribeiro, R., Batista, F.: Semantic search of mobile applications using word embeddings (2021)
18. Asmala, S.: Improving the search of medical articles using sentence transformers. In: HealthTech Proceedings, vol. 3, no. 1 (2023)
19. Pérez, J.C.: Chatbot/asistente virtual para mejorar la productividad de las tiendas (2024)
20. Bernal, T.A.: Generative ai: Fighting for child nutrition (2024)
21. Meta, React–a javascript library for building user interfaces. https://reactjs.org/, En línea. accedido el 4 de julio de 2025
22. Selenium, Selenium. https://www.selenium.dev/. En línea, accedido el 1 de agosto de 2025
23. Roy, D.: Django REST Framework: Building Web APIs with Python & Django. Packt Publishing (2020)
24. Aumüller, Y., Bernhardsson, E., Faithfull, A.: Ann-benchmarks: a benchmarking tool for approximate nearest neighbor algorithms. Inf. Syst. **87**, 101–112 (2020)
25. Mari and Eila.: The impact of maintainability on component-based software systems. In: 2003 Proceedings 29th Euromicro Conference, pp. 25–32. IEEE (2003)

Automatic Debate Generation for Spanish Educational Podcasts Using LLMs

Ricardo Lazo Vasquez(✉) and José Ochoa-Luna

Department of Computing, Universidad Católica San Pablo, Arequipa, Peru
{ricardo.lazo,jeochoa}@ucsp.edu.pe

Abstract. Online education is increasingly popular, but one challenge many teachers face is making their lectures engaging and interactive. Structured oral debates can be an effective way to achieve this, encouraging critical thinking and learning among students. However, scalable and reproducible methodologies for automatically generating such debate—particularly in Spanish—remain limited. This paper presents a modular approach for generating podcast-style educational debates in Spanish utilizing Large Language Models (LLMs). The proposed pipeline combines academic text summarization, structured prompting, and multilingual adaptation to produce topic-grounded, pedagogically meaningful conversations. The methodology is validated through a multidisciplinary case study that spans five academic domains and is evaluated using BLEU scores across various LLM configurations, including zero-shot and fine-tuned open-source models. Results show that high-quality, coherent debates can be generated even in low-resource settings, supporting the development of inclusive, AI-powered educational tools for Spanish-speaking learners.

Keywords: LLMs · Spanish NLP · Automated podcast generation · Online education

1 Introduction

Recent advances in Large Language Models (LLMs), such as Deepseek R1 and ChatGPT-4o, have demonstrated remarkable capabilities in text understanding, reasoning, and generation. These models present promising opportunities to transform digital education by enabling dynamic, personalized, and scalable content creation. Their ability to tailor outputs to individual learner needs opens the door to new instructional paradigms aligned with diverse learning styles—visual, auditory, reading/writing, and kinesthetic.

The global shift toward online education, accelerated by disruptions such as the COVID-19 pandemic, has revealed both the potential and limitations of current e-learning paradigms. Although access to asynchronous materials has increased, learners often continue to face challenges related to engagement, comprehension, and motivation. These limitations underscore the need for more interactive, contextualized, and learner-centered approaches.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 374–386, 2026.
https://doi.org/10.1007/978-3-032-20322-9_26

One of the most effective strategies for fostering active learning is structured debate. Rooted in Vygotskiĭ's sociocultural theory of learning, debate interaction plays a central role in the development of higher-order thinking by scaffolding learner understanding through guided exchanges [21]. Educational debate, in particular, allows learners to engage critically with content, articulate contrasting viewpoints, and internalize knowledge through collaborative reasoning.

In this work, we explore the automated generation of structured educational debates in a podcast format, a modality that supports auditory learners and enhances cognitive engagement. While tools such as Google's NotebookLM [14] have demonstrated the feasibility of generating podcast-style content grounded in user documents, such systems are closed-source and lack API access.

Moreover, most existing LLM-based educational pipelines are designed for English-language content and rely on freeform dialogue or static question answering. This creates a methodological gap in the generation of structured, pedagogically aligned debates in other languages. Spanish, despite being the native language of over 500 million people, remains underrepresented in multilingual educational NLP applications.

To address this gap, we propose a general, modular, and reproducible methodology to generate podcast-style educational debates in Spanish using LLMs. Our approach combines academic document summarization, structured prompting, multilingual adaptation, and model-agnostic debate generation. The resulting pipeline enables personalized, topic-specific content that can be embedded into Learning platforms for on-demand podcast creation.

We validate the methodology through a multidisciplinary case study using academic texts from five university-level domains: Artificial Intelligence, Biology, Computer Science, Economics, and Physics. Outputs are benchmarked against reference debates using BLEU scores, comparing both zero-shot and fine-tuned open-source models. Results show that even compact LLMs, when guided by a structured pipeline, can produce coherent, high-quality educational debates in Spanish.

This work contributes a practical and scalable framework for educational debate generation. By grounding the methodology in pedagogical theory and multilingual NLP, it bridges the gap between advanced language generation techniques and the instructional needs of Spanish-speaking learners. The methodology is compatible with both low-resource and high-performance environments, offering a replicable foundation for inclusive, AI-enhanced learning systems in a local environment as a starting point to deploy future solutions.

We also start evaluation in EduAI, a beta educational platform based on educational debates and tutoring that integrates the pipeline in their baseline solutions alongside other educational technologies for teaching and learning.

The remainder of this paper is organized as follows: Sect. 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the experimental setup. Section 5 reports and analyzes the results. Section 6 concludes the paper. Appendix A provides implementation details to support reproducibility and adaptation.

2 Related Work

Artificial Intelligence (AI) has significantly influenced educational landscapes through advancements in Natural Language Processing (NLP) and Large Language Models (LLMs). These systems are increasingly embedded in learning platforms through tools such as ChatGPT, Gemini, and NotebookLM [3,14], enabling document-grounded interaction, context-aware responses, and dynamic content synthesis. NotebookLM, in particular, allows users to query their own documents and receive contextualized explanations, though it remains limited by its closed-source nature.

The pedagogical relevance of debates as a learning tool is well established in educational theory. Vygotskiĭ's sociocultural framework positions debate and interaction as fundamental to cognitive development, particularly when learners are exposed to structured exchanges that scaffold understanding [21]. This theoretical foundation motivates the integration of structured conversational formats, such as educational debates, into AI-driven learning environments.

Despite the growing adoption of LLMs in education, research on the structured generation of pedagogical debates, especially in debate-style formats and non-English languages, remains limited. Vallecillo-Rodríguez et al. [17] proposed the use of LLMs to generate Spanish counter-speech for misinformation, illustrating the challenges and importance of adapting generative models for structured, responsible discourse in low-resource linguistic contexts.

Complementary work has examined the role of AI-generated feedback in learning. Ruwe and Mayweg-Paus [18] compared feedback from human instructors and AI systems on student argumentative writing, finding that learners often perceived the AI feedback as more credible and useful, suggesting a promising role for LLMs in metacognitive development through debate-based support.

On the generative side, Hu et al. [7] introduced the modular framework AMERICANO for argument generation using agent-based interaction and discourse decomposition. Although focused on English corpora, their architecture provides a conceptual foundation for scalable, multilingual educational debate generation.

Other alternatives, such as Tree-of-Debate [10], proposed the complete evaluation flow with a tree-based alternative to generate multi-persona debates about scientific papers. This alternative was evaluated against early stages of itself based on human-level tagging of each generated debate.

Similarly, other approaches such as MoonCast [9] aim for a deeper evaluation considering long-context debates. The alternative offers a full end-to-end approach from the prompt to audio generation based on English and the Chinese languages.

A full judge-based alternative was built by Liang et al. [12]. The authors proposed Debatrix, a graphical tool to evaluate LLMs on debates, looking for a winner and context-based metrics. This framework aims for self-evaluations against two LLMs. Similarly, Zhao et al. proposed Auto-Arena [23], looking to improve evaluations with 92.14% correlation with human evaluation.

Nonetheless, a gap persists in the availability of scalable methodologies for generating spoken educational content, particularly podcast-style debates grounded in academic texts and tailored for Spanish-speaking learners. The framework proposed in this paper addresses this gap through a modular, multilingual pipeline for generating structured educational debates using LLMs.

3 Proposed Methodology for Spanish Educational Debate Generation

This section outlines a modular and reproducible methodology for generating structured educational debates in podcast format, specifically designed for Spanish-speaking learners. The proposed system serves as the core pipeline of EduAI, a prototype educational platform that enables on-demand generation of topic-specific, debate-style podcasts from academic texts. The methodology emphasizes model agnosticism, educational alignment, and scalability, and is illustrated in Fig. 1.

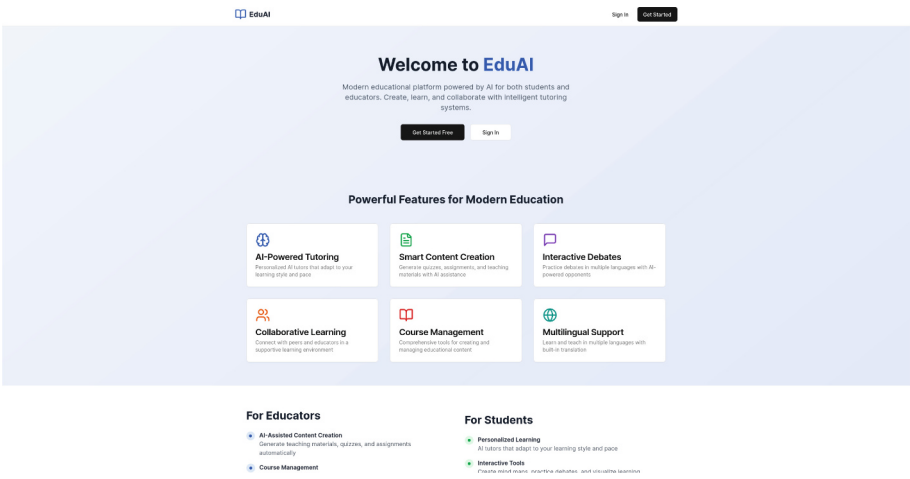


Fig. 1. Proposed methodology integrated in EduAI, a prototype platform for tutoring.

EduAI aims to be tested in future iterations of the proposal with the goal of being a platform for tutoring for teachers and students. On the other hand, the full proposal methodology is represented in the Fig. 2 and broken down in the following subsections.

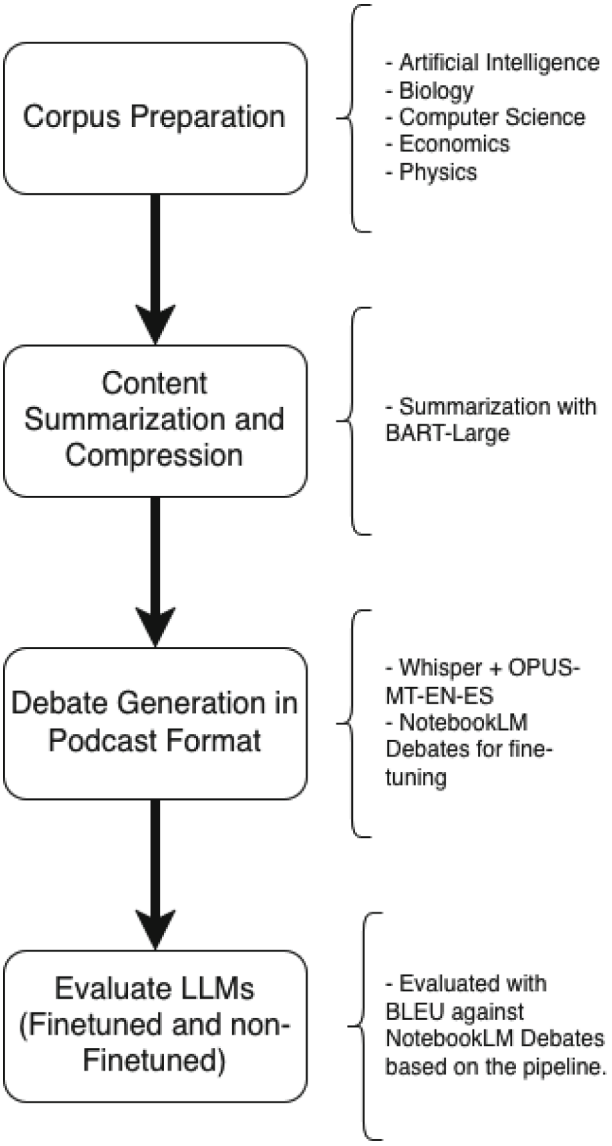


Fig. 2. Pipeline of the proposed methodology.

3.1 Corpus Preparation

We began by curating a domain-balanced corpus of 50 open-access academic papers spanning five disciplines: Artificial Intelligence, Biology, Computer Science, Economics, and Physics. This mirrors the typical distribution of materials encountered in university-level coursework. All documents were in English and selected based on accessibility, clarity, and disciplinary relevance.

3.2 Content Summarization and Compression

To prepare the documents for generation, we employed a page-level summarization pipeline using BART-large [11]. The summaries were then recursively compressed to ensure they fit within the context window of the target LLMs. A fixed instructional prompt guided the model to prioritize pedagogically significant content while avoiding non-informative artifacts. The resulting text serves as the semantic scaffold for downstream debate generation.

3.3 Debate Generation in Podcast Format

The summarized content was transformed into two-speaker, podcast-style debates using structured prompting. A consistent Spanish-language prompt was used across all experiments:

```
Usando este resumen: {summary}
Genera un debate tipo podcast entre dos personas en español.
```

The conversations are generated in a single pass (i.e., not turn-by-turn), simulating academic conversations or friendly debates that emphasize contrastive reasoning and conceptual clarification. This format draws on sociocultural learning theory, particularly the role of debate in scaffolding higher-order thinking.

Initial reference debates were created using Google NotebookLM, transcribed via Whisper [16], and translated to Spanish using OPUS-MT EN-ES [20], both LLM-based applications that can integrate into the code pipeline. These transcripts serve as evaluation references and fine-tuning targets.

3.4 Model Configurations and Modularity

The pipeline supports multiple LLM configurations:

- **Zero-shot inference:** Using open-source models executed via Ollama.
- **Fine-tuned generation:** Via LoRA adapters on compact models (e.g., Gemma 2B, Phi 1.5B, Qwen 1.5).
- **API-based benchmarking:** Using Gemini Flash 2.0 to establish an upper-bound quality reference.

Each pipeline stage—summarization, translation, generation—can be replaced or upgraded independently. This modularity supports flexible deployment across constrained or high-resource environments. Implementation details (including code, parameters, and training configurations) are provided in Appendix A.

3.5 Evaluation Strategy

BLEU scores were used as an automated metric to compare the generated debates to those produced by NotebookLM. Although BLEU is limited in capturing semantic nuance or pedagogical effectiveness, it provides a reproducible measure of lexical alignment and language fluency across models and settings. Detailed results are reported in Sect. 5.

3.6 Educational Integration and Use Cases

The system is designed for personalized integration within learning platforms. A typical use case involves a learner selecting a topic or uploading material, receiving a podcast-style debate as output. This enables auditory reinforcement of complex concepts without the need for real-time instruction.

While the full EduAI system is under development, the key pipeline components—summarization, prompting, generation, and evaluation—are fully reproducible and will be shared via a public GitHub repository.

3.7 Limitations and Future Directions

This first version of the methodology omits several advanced features, such as bidirectional learner feedback, adaptive personalization, and TTS for native Spanish audio. While technically feasible, these remain out of scope for the current evaluation.

The framework, however, is designed to accommodate such extensions. Its flexibility and language-aware design enable future expansions into other languages, disciplines, and interaction modalities, contributing toward scalable, inclusive educational tools powered by LLMs.

4 Case Study: Applying the Methodology to Multidisciplinary Academic Content

To validate the proposed methodology, we conducted a case study using academic content from five university-level disciplines: Artificial Intelligence, Biology, Computer Science, Economics, and Physics. This selection reflects the variety of topics typically encountered in higher education and allows for evaluating the pipeline's generalizability across domains.

4.1 Corpus Design and Language Constraints

We assembled a 50-document corpus composed of open-access academic texts, equally distributed across the five disciplines. All documents were originally written in English, mirroring the dominant language in academic publishing. The corpus was selected through domain-aware retrieval queries and was divided into training and test splits (80/20) to support both inference and fine-tuning stages.

4.2 Reference Debate Generation

To establish a benchmark for evaluation, we used a fully automated multilingual adaptation pipeline based on three state-of-the-art tools. Source documents were summarized and passed through Google's NotebookLM for structured educational debates synthesis. We generated transcripts by transcribing the English audio with Whisper and translating them using OPUS-MT EN-ES. These Spanish conversations served both as reference outputs and fine-tuning targets. No manual curation was applied, preserving automation and reproducibility.

4.3 Model Evaluation Scope

We evaluated the methodology under three deployment modes:

- **Zero-shot inference**, using open-source models like Phi [1,2], Gemma [19], Qwen [4], Mistral [8,15], TinyLlama [22], and DeepSeek [5,13].
- **Fine-tuning**, applying LoRA [6] to compact multilingual models (e.g., Phi 1.5B [2], Qwen 1.5B [4], Gemma 2B [19]) using the Spanish debates as training data.
- **Proprietary benchmark**, comparing results to Google's Gemini 2.0 Flash API.

The BLEU score was used as a proxy for linguistic similarity and structure alignment, comparing model-generated debates with the NotebookLM-based Spanish references. Full implementation details are provided in the appendix.

4.4 Deployment and Reproducibility

Experiments were run in a local environment equipped with a consumer-grade GPU. The methodology, prompts, and models used are fully documented and open to replication using different model configurations or languages, reflecting the modular design of the pipeline.

5 Results

We evaluated the performance of the proposed methodology across a diverse range of Large Language Models (LLMs), testing three deployment modes: zero-shot inference, fine-tuning with LoRA adapters, and proprietary APIs. The goal was to assess the pipeline's ability to generate coherent, pedagogically relevant debate in Spanish from academic texts.

5.1 Evaluation Metric

To measure linguistic alignment between model outputs and reference debates, we used the BLEU metric (via SacreBLEU), normalized to a 0âĂŞ100 scale. While BLEU is traditionally applied to machine translation, it offers a consistent basis for comparing:

- Surface-level similarity to benchmark debates.
- Structural consistency across multilingual outputs.

Given the creative nature of educational debate generation, we interpret BLEU as a *relative*—not absolute—indicator of quality, suitable for comparing performance under controlled conditions.

5.2 Model Categories and Scores

Table 1 reports BLEU scores across model categories. The best result per group is shaded (green for open-source, blue for overall), and the top scores per column are underlined.

Table 1. BLEU scores (0âĂŞ100 scale, computed with SacreBLEU) for generated podcast-style debates. Best per row in bold; best per language (column) underlined; best per model group in light green; overall best model in light blue.

Vanila Models (Using Ollama)		
Model	BLEU Spanish	BLEU English
deepseek-llm	0.000130	**0.028045**
deepseek-v2	**0.024395**	0.019025
gemma	0.000030	**0.000914**
gemma2	0.007451	**0.012026**
llama2	**0.048781**	0.005074
llama3	**0.009224**	0.004870
mistral	0.004241	**0.012830**
mistral-nemo	0.002823	**0.034916**
mistral-small	**0.121382**	0.084437
mistrallite	0.000000	0.000000
phi	0.000000	0.000000
phi3	**0.139620**	0.063222
phi3.5	0.082688	**0.090814**
phi4	**0.055309**	0.032736
phi4-mini	**0.403415**	0.025366
qwen	0.000000	0.000000
qwen2	0.007180	**0.017294**
qwen2.5	0.046038	**0.111163**
tinyllama	0.000129	**0.000825**
Fine-tuned Models		
Model	BLEU Spanish	BLEU English
Gemma-2b	**0.323689**	0.296376
Phi-1.5	**0.007208**	0.000000
Qwen1.5 0.5B	**0.148865**	0.083888
Qwen1.5 1.8B	0.206677	**0.457367**
API tested		
Model	BLEU Spanish	BLEU English
Gemini 2.0 Flash	**0.752201**	0.546049

5.3 Zero-Shot Results

Among vanilla models, **Phi 4-mini** achieved the highest BLEU score in Spanish (0.4034), suggesting that even compact models can yield quality outputs when guided by structured prompting. **Phi 3.5** and **Mistral-Small** also performed well in both Spanish and English, reinforcing the pipeline's value for lightweight deployment.

In contrast, models like **TinyLlama**, **Qwen 1.0**, and earlier versions of **Gemma** failed to generate meaningful outputs under zero-shot settings, confirming the importance of architectural scale and language exposure.

5.4 Fine-Tuned Model Performance

Fine-tuning with Spanish podcast debates significantly improved results. **Gemma 2B** and **Qwen 1.5 1.8B** outperformed their vanilla counterparts, with the latter reaching 0.4574 BLEU in English—the best score among all open models. These results validate the utility of the pipeline not just for inference, but also as a high-quality data source for domain-specific adaptation.

5.5 API Benchmarking

Gemini 2.0 Flash, tested as an upper-bound reference, yielded the top BLEU scores overall. While this confirms the strength of proprietary models, the gap between them and fine-tuned open-source alternatives is narrowing, especially for Spanish generation tasks. This demonstrates the pipeline's effectiveness in low-resource, privacy-conscious, or infrastructure-limited settings.

5.6 Insights and Observations

The variability in BLEU scores reflects the open-ended nature of the task. Unlike translation, educational debate generation allows for a broad range of valid responses. Still, structured prompting and summarization helped reduce this variance across models.

Interestingly, several models performed better in Spanish than English—despite training predominantly in English—likely due to the pipeline's Spanish-centric design. This suggests that targeted prompting and summarization strategies can offset the imbalance in multilingual training data.

6 Conclusions

We presented a modular and replicable methodology for generating podcast-style educational debates in Spanish using Large Language Models (LLMs). By integrating summarization, structured prompting, and multilingual adaptation, the pipeline enables the automated creation of topic-grounded, pedagogically meaningful debates that can be embedded into personalized online learning environments.

The methodology was validated through a multidisciplinary case study covering five academic domains and tested under both zero-shot and fine-tuned configurations. Evaluation with BLEU scores demonstrated that even compact, open-source models—when guided by a structured pipeline can generate coherent and relevant educational debates in Spanish. Moreover, the same outputs proved suitable as high-quality training material for fine-tuning lightweight models, broadening the practical utility of the framework.

Rather than introducing a new model, this work addresses a key methodological gap: the lack of reproducible strategies for educational debate-generation in underrepresented languages. It offers a technically feasible and pedagogically grounded solution for scalable content generation, adaptable to various educational contexts and compatible with a wide range of language models.

The results confirm the robustness and applicability of the proposed approach. Its successful deployment across diverse models and academic domains underscores its potential as a foundation for inclusive, AI-enhanced learning systems, particularly in multilingual and low-resource environments.

In future works, it is possible to increase the model's validation severity using MoonCast, Auto-Arena, or Debatrix. Also, it is possible to make the models compete against each other using these approaches.

A Implementation Details

A.1 Environment and Dependencies

All experiments were conducted on a local Ubuntu 24.04 workstation equipped with an NVIDIA RTX 4070 Ti SUPER GPU. The following core libraries and tools were used:

- `transformers` for model loading and inference.
- `accelerate`, `peft`, and `trl` for LoRA-based fine-tuning.
- `sacrebleu` for evaluation metrics.
- `sentencepiece`, OpenAI's `whisper`, and `opus-mt-en-es` for tokenization and multilingual preprocessing.

A.2 Prompt Engineering

Structured prompting was applied to guide debate generation. The following templates were used:

Spanish Prompt:

```
Usando este resumen:
{summary}
Genera un debate tipo podcast entre dos personas en espa\~{n}ol.
```

English Prompt:

```
Based on this summary:
{summary}
Generate a podcast debate between two people.
```

The entire debate was generated in a single forward pass to simulate a natural two-speaker podcast. No turn-by-turn generation or user interaction was involved.

A.3 Dataset Format

Each training instance consisted of a JSON object containing:

- `summary`: the academic summary used as input.
- `debate`: the generated podcast-style debate (target output).

All data was tokenized using the appropriate model tokenizer, with a maximum context length of 2048 tokens.

A.4 Training Setup

Fine-tuning was performed using LoRA with the following hyperparameters:

- Rank: 64
- Alpha: 32
- Dropout: 0.1
- Epochs: 10
- Batch size: 2
- Learning rate: 5×10^{-5}

Evaluation was conducted at the end of each epoch using BLEU scores against the NotebookLM-derived reference transcripts.

A.5 Code and Reproducibility

All implementation resources of our experiments are available via a public GitHub repository https://github.com/TheReverseWasp/Debate-Generation-Spanish. The repository includes:

- Notebooks for summarization, structured prompting, and fine-tuning.
- Sample training and evaluation data in JSON format.
- Scripts for batch BLEU evaluation and corpus preprocessing.

A.6 Pipeline Application

The pipeline was created looking to be applied in an innovative educational platform for students and teachers, that can create their own podcasts, debates, and tutoring alongside lessons and tests **EduAI**. We aim to test this platform in future iterations of the research.

References

1. Abdin, M., et al.: Phi-3 technical report: a highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Abdin, M., et al.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)
3. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
4. Bai, J., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Bi, X., et al.: Deepseek llm: scaling open-source language models with longtermism (2024)
6. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. In: ICLR, vol. 1, no. 2, p. 3 (2022)
7. Hu, Z., Chan, H.P., Yin, Y.: Americano: argument generation with discourse-driven decomposition and agent interaction. arXiv preprint arXiv:2310.20352 (2023)
8. Jiang, A.Q., et al: Mistral 7b (2023). https://arxiv.org/abs/2310.06825
9. Ju, Z., et al.: Mooncast: high-quality zero-shot podcast generation. arXiv preprint arXiv:2503.14345 (2025)
10. Kargupta, P., Agarwal, I., August, T., Han, J.: Tree-of-debate: multi-persona debate trees elicit critical thinking for scientific comparative analysis. arXiv preprint arXiv:2502.14767 (2025)
11. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
12. Liang, J., et al.: Debatrix: multi-dimensional debate judge with iterative chronological analysis based on llm. arXiv preprint arXiv:2403.08010 (2024)
13. Liu, A., et al.: Deepseek-v2: a strong, economical, and efficient mixture-of-experts language model (2024)
14. Martin, R.: Introducing NotebookLM (2024). https://blog.google/technology/ai/notebooklm-google-ai/
15. Mistral AI: Mistral nemo (2024). https://mistral.ai/news/mistral-nemo. accedido el 14 de mayo de 2025
16. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
17. Rodríguez, M.E.V.: Adaptation of large language models for Spanish text generation in responsible AI problems (2024)
18. Ruwe, T., Mayweg-Paus, E.: "your argumentation is good", says the ai vs humans-the role of feedback providers and personalised language for feedback effectiveness. Comput. Educ. Artif. Intell. **5**, 100189 (2023)
19. Team, G., et al.: Gemma: open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
20. Tiedemann, J., Thottingal, S.: Opus-mt–building open translation services for the world. In: Annual Conference of the European Association for Machine Translation, pp. 479–480. European Association for Machine Translation (2020)
21. Vygotskiĭ, L.S., Cole, M.: Mind in Society: The Development of Higher Psychological Processes (1978)
22. Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: an open-source small language model (2024). https://arxiv.org/abs/2401.02385
23. Zhao, R., Zhang, W., Chia, Y.K., Xu, W., Zhao, D., Bing, L.: Auto-arena: automating llm evaluations with agent peer battles and committee discussions. arXiv preprint arXiv:2405.20267 (2024)

When One Size Doesn't Fit All: A Study on Intrinsic Iterative Self-correction in LLMs for Hallucination Control

Svetlana Esina(✉) and Diana Inkpen

University of Ottawa, Ottawa, ON K1N 6N5, Canada
sesin024@uottawa.ca
https://www.uottawa.ca/en

Abstract. Hallucinations refer to the tendency in Large Language Models to generate content that is not supported by external sources. Their presence reduces the reliability of the model and spreads misinformation. This study deep-dives into a strategy that uses a model's inherent knowledge to combat hallucinations, called intrinsic self-correction. We apply self-correction multiple times in an extractive question-answering setting on multiple models of different families and sizes. Our findings suggest that the approach successfully mitigates hallucinations in the Qwen and Gemma families, but is not completely transferrable to others like Llama. A one-size-fits-all approach to self-correction is therefore, ineffective, requiring tailoring to a model's specific reflective and instruction-following abilities to prepare for the adoption of this strategy into the real world.

Keywords: Large Language Models · Natural Language Processing · Self-correction · Question-Answering · Hallucinations

1 Introduction

Although the use of Artificial Intelligence (AI) technologies and Large Language Models (LLMs) has become more widespread in the past century, many still retain doubts about the reliability of these models [5,14,21,22]. Especially in the task of question-answering (QA), models have a tendency to respond with information that is not supported by any external source [15]. These answers appear to be coherent and grammatically correct, ultimately tricking unsuspecting individuals into believing they are truthful. This behaviour is commonly referred to as a hallucination [5,6]; it is said that the model internally hallucinates some relationship between information that was acquired during its pre-training, causing unsupported statements to appear in its responses. Thus, without any prior knowledge, it becomes almost impossible to distinguish a hallucination from a truthful response. These hallucinations inspire the spread of misinformation [5] or, otherwise, undermine the performance of the model on a task. This behaviour becomes particularly critical when applied in domains such

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 387–402, 2026.
https://doi.org/10.1007/978-3-032-20322-9_27

as medical or legal, where responses provided by these LLMs directly impact decisions that may alter one's life or livelihood [5,6].

One way to mitigate hallucinations in LLMs is to apply a post-generation strategy called self-correction. Self-correction, also known as self-reflection, attempts to mimic how humans learn; it starts by creating an initial draft for the response to a question, then evaluates and critiques this draft, and finally refines it by incorporating the feedback from the previous step, to create the final version of the response [16]. Particularly with LLMs, self-correction is implemented through prompt engineering, where the role the model plays is defined in the instructions passed to the model with the aid of few-shot examples [2,13]. This strategy eliminates the need to have a human in the loop, any dependence on external models and the requirement for additional training, making it a fairly low-cost solution. However, research done in this field obtained mixed results on the effectiveness of self-correction for hallucination mitigation; it is unclear when the strategy works or in what cases it should be applied. This uncertainty has prompted the exploration of more rigorous methods, leading to the introduction of iterative self-correction. By moving beyond a single revision, this approach allows the model to progressively refine its response over several cycles, potentially mitigating the inconsistencies found in a single cycle of self-correction and leading to a more reliable final output [3,13,16]. Although iterative self-correction has been explored in a variety of tasks and different configurations, it has yet to be extensively studied in the context of extractive QA - a foundational task for factual retrieval and the recognition of knowledge boundaries.

This study investigates the effectiveness of a specific implementation of iterative self-correction in extractive QA, where the answers to the proposed questions are found within the context of the prompt. We evaluate its performance on a base Qwen2.5-1.5B model, and contrast it with the performance on models with different sizes (0.5B and 3B) and different families (Gemma3 and Llama3.2) to allow for a more insightful and less biased analysis. Specifically, this study investigates the following research questions: is the implementation effectively performing iterative self-correction on the base model (RQ1), what is the effect of iterative self-correction using the base model on hallucinations (RQ2), can the same implementation be used on models of different size and families to achieve the same effect on hallucinations (RQ3). To address these questions, the study conducts a series of controlled experiments where responses from every iteration are recorded and later analyzed through metrics like the exact match (EM) rate, the METEOR score, the BERT F1-score and the SummaC score. As such, this research aims to shed light on the conditions under which iterative self-correction is most effective for extractive QA and to determine its generalizability across a range of model architectures and scales to prepare for its potential adoption in real-world settings.

2 Related Work

Strategies to mitigate hallucinations through self-correction are mostly split into two categories: one that provides external knowledge to the model to ground

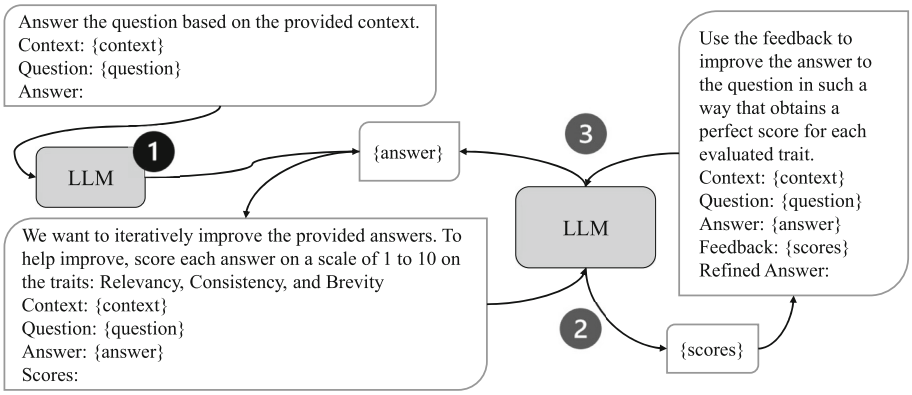
itself to a truth, and one that does not. Self-correction based on external knowledge relies on an abundance of extra information in the appropriate domain to be integrated in each iteration and, thus, presents an inherent limitation to this strategy's employment in real-world contexts, where extra information may not always be available. In contrast, intrinsic self-correction relies on the assumption that a model can use its inherent knowledge to identify what is wrong with a response, though its effectiveness as a hallucination mitigation strategy has yet to be confirmed [7,9]. Some studies achieve significant improvements [13,20]; however, a common issue with their experiments is the use of oracle information in the reflection process or as a stopping criteria [4,7,9]. Mixed results are obtained when testing self-correction with fair and realistic evaluation on reasoning tasks [7,9], relying on a model's comprehension of the question, the question difficulty, and the thoroughness of the generated feedback to achieve an significant improvement.

In the field of QA, a common task being explored with iterative self-correction is the multiple-choice [12,13], where the question suggests options for the model to select from. Although this tasks allows for more straightforward evaluation, as the correct answer is selected from a finite set of options, it does not fully capture the real-world application of QA in LLMs where these options are often not available. As such, for this study, we look at extractive QA - a QA task where the prompt does not incorporate a set of answers for the LLM to pick from but requires the model to use the information from context provided to answer the question. This task closely mirrors the real-world setting, where the model would have seen the answer during pre-training or in the prompt (e.g., in retrieval-augmented generation), or not at all. To sum up, this study incorporates several novel considerations, from witholding oracle information in the reflection or refinement steps, to testing on a more realistic task (i.e., extractive QA), to exploring a wider range of models from those typically found in related works (i.e., Qwen, Gemma, and Llama) in order to determine the effectiveness of iterative self-correction for hallucination mitigation.

3 Methodology

Evaluating the effectiveness of iterative self-correction involves comparing the answers generated at iteration 0 (when no self-correction is applied), and at iteration n (when self-correction is applied n times, where $n > 0$) as shown by Fig. 1. For iteration 0, the LLM generates a response given only the question and context. All following iterations rely on the previously generated response and employ a 2-step process: the LLM critiques it using a rubric, and then generates a new refined response based on this feedback. The critiquing model evaluates each response using relevancy (does the answer address the question), consistency (is the answer supported by the context), and brevity (is the answer concise and without any unnecessary details), and requests the model to reflect on the given traits through both descriptive feedback and quantitative feedback (i.e. scoring each trait out of 10). This feedback strategy should allow the model to

provide more nuanced and insightful feedback [16], thereby enabling it to perform precise refinements to its response in subsequent iterations. The exact prompt used in the experiments was based of [13]. In addition to the main iterative self-correction experiment, a self-consistency baseline is established for iteration 0. It involves generating the initial response to the prompt 10 independent times, by varying the seed, and retaining the generated answer that is most common to help evaluate the quality of the model's initial, uncorrected output, and allow this study to quantify the significance of any increase or decrease in performance due to self-correction.



Algorithm 1 Iterative self-correction

Require: input x, model M, prompts $p_{generate}$, $p_{critique}$, p_{refine}, max iterations N

$y_0 \leftarrow M(p_{generate}||x)$

for iteration $n = 1$ to N **do**

Generate feedback $f_{n-1} \leftarrow M(p_{critique}||x||y_{n-1})$

$y_n \leftarrow M(p_{refine}||x||y_{n-1}||f_{n-1})$

end for

return y_N

Fig. 1. Pipeline and algorithm for iterative self-correction, where first the initial answers are generated (step 1) and then iteratively critiqued (step 2) and refined (step 3).

3.1 Large Language Models

The models attempted in this study were chosen based on the specific criteria suggested by [13]. The base model was selected to be the Qwen2.5-1.5B [24,26] decoder-only model, while Qwen2.5-0.5B [24,26] and Qwen2.5-3B [24,26] are used to test the strategy on models of different size, and Gemma3-1B [23] and Llama3.2-1B [25] to test on models from different families. All of the models selected are instruction-tuned to be able to adequately follow the instructions

provided through the prompt and, thus, cater the output of the models to the expected format (when adhering to the critic or refiner roles). Additionally, due to computational resource constraints, quantized versions of the models are used. For the Qwen family, models employing the Activation-aware Weight Quantization (AWQ) method were selected, allowing for the 4-bit quantization of the model to speed up inference without significantly impacting performance [10]. For the Llama and Gemma families, the models were set to be 4-bit quantized during the experiments using the BitsAndBytes library. The use of quantization across all models ensures a consistent computational environment, allowing for a fair comparison of the proposed methodology.

3.2 Evaluation

One part of the evaluation process is concerned with the model's inherent ability to adopt the critic and refiner roles through the prompt. The process involves comparing the model's behaviour in the experiment with the following expectations: A) there is a change in the new refined response when the combined feedback score is not perfect, and there is no change when it is (30 out of 30), B) there is an upward trend in the feedback scores generated with every subsequent iteration of self-correction, and C) there is a downward trend where less changes are made to the response with every subsequent iteration. These behaviours allow us to analyze how well the iterative self-correction strategy is being implemented (RQ1). The other part of the evaluation process is concerned with the hallucination rate. We define hallucinations to be all answers that are not similar to any of the provided correct answers and employ multiple metrics to assess this similarity after every iteration. The EM rate is used to assess how many times the generated response matches an expected one. However, this metric does not take into account synonyms and paraphrasing that can also be used to correctly answer a question. The METEOR score computes the number of generalized unigram matches (i.e., surface forms, stemmed forms, meanings) between the predicted and referenced strings [1], allowing for a more in-depth analysis of unigram overlap compared to EM rate. The BERT score evaluates the semantic similarity of each token in the generated and expected strings through pre-trained contextual embeddings [27]. Only the F1-score is retained for the BERT score as it combines metrics for relevance and completeness of the response - both of which are required in extractive QA. Lastly, the SummaC score leverages Natural Language Inference (NLI) models to measure factual consistency between sentences in the predicted and referenced strings [8], quantifying the truth-alignment of the generated responses. A SummaC score close to 1 indicates a higher factual consistency. By using multiple metrics, this study aims to incorporate multiple perspectives on the hallucinations generated during self-correction (RQ2 and RQ3).

4 Experiments

4.1 Datasets

Two datasets are selected for the QA task. The first dataset, SQUAD 2.0 [18,19], includes questions posed by crowdworkers on a set of Wikipedia articles and their answers. It contains QA pairs from a wide variety of domains - from law to health to history - allowing experiments to examine the self-correction strategy in multiple spheres of knowledge. This dataset came pre-split into train (130,319 records) and validation (11,873 records) subsets, with all records including the context, the question and the list of correct answers. There were no empty questions and no leakage of records from the train to the validation subsets. To reduce the time required to run the experiments, the scope of the study was also limited to the first 5,000 examples from the validation split. 51% of these examples contained no answer based on the context provided, and thus were replaced with "Cannot be determined". The train subset was exclusively reserved for few-shot examples to supplement the validation records with. The second dataset, TruthfulQA [11,17], includes human-crafted QA pairs on online articles (including Wikipedia) that were designed to elicit imitative falsehoods. This dataset also covers a variety of domains, ensuring diversity in the types of topics being tested. The original dataset consists of 817 records, where each record includes the source url, the context, the question, and a list of correct answers. However, after filtering out all records with contexts that were too long (according to the upper whisker of their box plot) and all records with contexts that were improperly retrieved from the source url, only 534 records remained. Additional pre-processing steps also shortened the context length for each record to be able to fit within the models' maximum input size. The method involved creating sliding windows (of length 4,000) from the full context and calculating the cosine similarity between the combination of the question and its correct answers and each window. The window with the highest cosine similarity score was then selected as the new context for each record. Further, any of the correct answers that said "I have no comment" or "The question is ambiguous", was replaced with "Cannot be determined". Finally, the dataset was divided into 450 records for the validation subset and 84 records for the train subset, where the latter was exclusively used for few-shot examples. While the first dataset contains more examples, the QA pairs are more focused on the retrieval of specific facts from the context, limiting the responses to single words or phrases. In contrast, the second dataset assesses the ability to provide detailed explanations and the use of more complex reasoning. As such, by incorporating both datasets, the study aims to provide a more comprehensive evaluation of the self-correction strategy in the QA setting.

4.2 Implementation of Pipelines

The experiments were conducted on a Kaggle Notebooks environment, with access to 2 T4 GPUs (with 15 GiB of memory each) and 29 GiB of RAM[1].

[1] Implementation code is available at https://github.com/Svetlana154/SelfRefl.

The total number of iterations for each experiment is 10, to give the models ample chances to recognize and correct any mistakes in the responses, and to be compared against the self-consistency baseline (SC_baseline) that also uses 10 generations. To avoid excessive overhead for iteration 0, the selection of few-shot examples from the training split was completed separately. GPT2 was used to obtain contextual embeddings of the examples in the training split and the top 2 (with the highest cosine similarity score) were selected as the few-shot examples for every entry in the validation split. Thus, the main experiments only involved generating the appropriate prompt template, loading in the model, and querying the model with the prompt. For iteration 0, the maximum number of new tokens the models could generate was set to 50, as the responses were expected to be brief. For iteration n, few-shot examples consisted of 3 human-labeled examples from the training split of the respective dataset, were kept static for the critiquing and refining prompts. Thus, the model could identify its role in the strategy and align itself with the expected format of the response. To stay consistent with the generation at iteration 0, the generation at the refinement step was also limited to 50 new tokens, while the generation for the feedback step permitted 200 new tokens (to allow for a more verbose reflection).

5 Results and Analysis

5.1 Base Model

Evaluation of Model Behaviour. Figure 2 reflects the metrics captured for the behaviour of the base Qwen2.5-1.5B model throughout the iterative self-correction experiment. Given a perfect feedback score, the refined response is expected to stay the same and the results of the experiment mostly conform to this behaviour as more iterations are processed, with the no-change rate being 89.97% at iteration 10 for SQUAD 2.0 and 95.96% at iteration 9 for TruthfulQA. However, initially, the rate starts much lower, suggesting that even though the critic considered the response to be perfect, the refiner still proceeded to change the response. The opposite is observed when the feedback score is not perfect; the model tends to keep the same response even if the feedback identifies a problem, suggesting that the model is not relying on its own quantitative feedback for its refinement. In terms of the average feedback score per iteration, the results obtained for the two datasets show no significant improvement with more iterations; the average combined and individual scores do not vary much; most perfect answers are not being changed and most non-perfect answers remain the same, causing the scores to appear stable as well. The response retention rate also confirms that over 85% of all answers are kept the same in iteration 4, and gradually increase with more iterations, which follows the increase in no-change rate given a perfect score with more iterations.

Evaluation with Ground Truth. To determine whether the retained responses are getting closer to the ground truth, we consider the change in EM rate, METEOR score, BERT F1 and SummaC scores compared to iteration 0

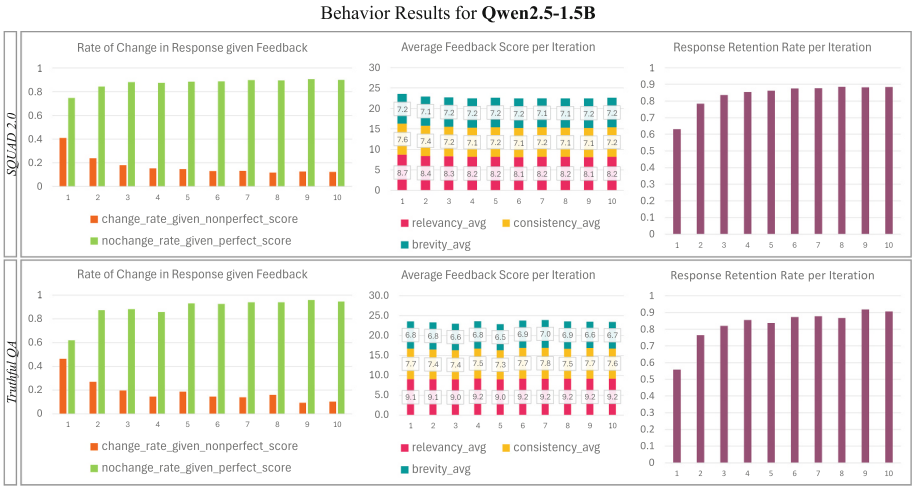


Fig. 2. The behaviour of the iterative self-correction experiment for Qwen2.5-1.5B on the SQUAD 2.0 and the TruthfulQA datasets.

(refer to Fig. 3). For the SQUAD 2.0 dataset, the results indicate a significant improvement in all metrics with a maximum increase of 3.32% for the EM rate (at iteration 2), 7.75% for the METEOR score (at iteration 4), 0.61% for the BERT F1-score (at iteration 2), and 3.92% for the SummaC score (at iteration 4). The iterative self-correction strategy enhances model performance by increasing the rate of verbatim matches with the expected answers, while also producing semantically closer outputs for responses. A similar trend in observed for the TruthfulQA dataset. Noticeably, the EM rate decreases with most iterations (except for iteration 6). Although this behaviour implies that more of the refined answers do not exactly match what is expected, the significant increase in METEOR score - a high of 8.91% at iteration 4 - suggests that the answers are still being improved using similar unigrams. The BERT F1 and SummaC scores back this idea up through the 1.34% (at iteration 5) and 1.20% (at iteration 5) increases, where both exceed their self-consistency baselines.

5.2 Qwen Family

Qwen2.5-0.5B. In reference to the larger 1.5B model, the 0.5B model produces very similar results on the SQUAD 2.0 dataset, but contrastingly different ones for TruthfulQA (as shown in Fig. 3). For SQUAD 2.0, all metrics show a significant improvement after applying self-correction: the EM rate reaches a high of 8.08% at iteration 7, the METEOR score - a high of 6.47% at iteration 7, the BERT F1-score - a high of 1.58% at iterations 7 and 9, and the SummaC score - high of 7.87% at iterations 7 and 9. The performance significantly improves from the first iteration, suggesting that the responses are getting closer to the ground truth, and that hallucinations are being mitigated. In the TruthfulQA

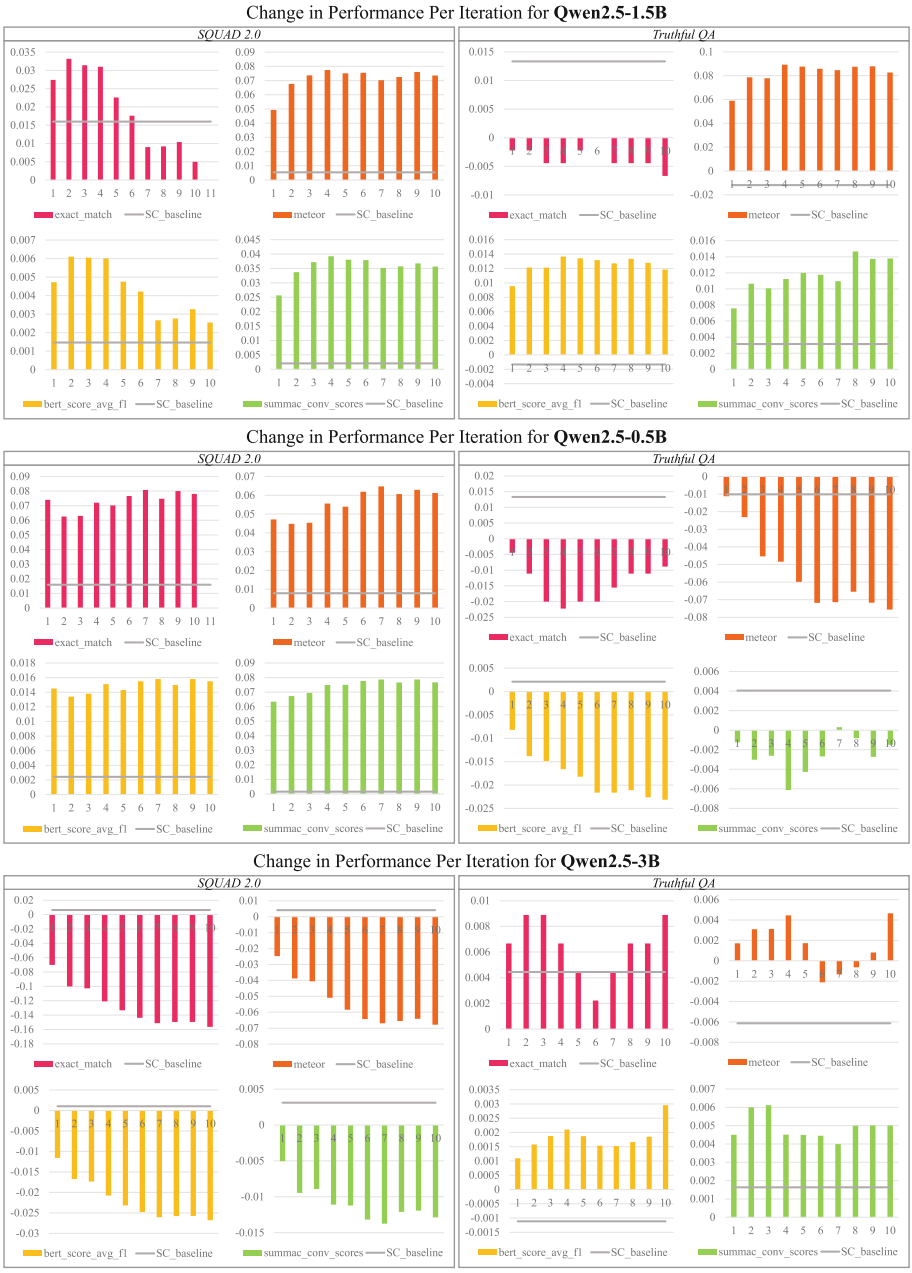


Fig. 3. Changes in the exact match rate (pink), the METEOR score (orange), the BERT F1-score (yellow), and the SummaC score (green) from iteration 0 to each subsequent iteration for the SQUAD 2.0 (left) and TruthfulQA (right) datasets, using Qwen2.5-1.5B (top), Qwen2.5-0.5B (middle), and Qwen2.5-3B (bottom). (Color figure online)

scenario, any iteration of self-correction generally hinders the performance of the smaller model. The EM, METEOR and BERT F1 all show a downward trend in performance as more iterations are applied. The SummaC score also generally decreases with iteration 7 being the only exception with an increase of 0.03%. No captured metric surpasses the self-consistency baseline in any iteration, suggesting that iterative self-correction was not effective on this dataset.

Qwen2.5-3B. The opposite effect to the smaller 0.5B model is observed in the results for the 3B model (Fig. 3); iterative self-correction hinders performance on the SQUAD 2.0 dataset, while improving performance for the TruthfulQA one. On SQUAD 2.0, all metrics show a decrease in performance starting with the first iteration. The EM rate suffers the most, with the minimum change in performance being −7% (at iteration 1) and the maximum being −15.64% (at iteration 10). Additionally, the results generated at iteration 0 are worse than the self-consistency baseline in terms of every metric, implying that most of the initial responses used in the self-correction experiment were further away from the ground truth than expected. For the TruthfulQA dataset, iterative self-correction improves the performance of the model, as shown by the increase in all evaluation metrics; the EM rate achieves a maximum increase of 0.89% (at iterations 2, 3, and 10), the METEOR score an increase of 0.47% (at iteration 10), the BERT F1-score an increase of 0.3% (at iteration 10), and the SummaC score an increase of 0.61% (at iteration 3). The improvement for all metrics is significant enough to exceed their self-consistency baseline from iteration 0.

5.3 Other Families

Gemma3-1B. The Gemma model generally improves with iterative self-correction on the SQUAD 2.0 dataset (Fig. 4). All metrics show a positive change in performance, achieving an increase of 3.76% in the EM rate (at iteration 3), an increase of 1.63% in the METEOR score (at iteration 4), an increase of 0.79% in the BERT F1-score (at iteration 4), and an increase of 2.40% in the SummaC score (at iteration 9). Additionally, all increases in performance are above the self-consistency baseline, showing the significance of the hallucination mitigation effect. In the TruthfulQA dataset, iterative self-correction produces better results for the BERT F1-score and the SummaC score at all 10 iterations, whereas the EM rate and METEOR score decreases for all 10 iterations (achieving the smallest decrease only at iteration 1). These observations suggest that for this dataset, the model generates answers that appear more semantically and contextually close to the ground truth, as opposed to using the exact same words or unigrams, explaining the discrepancy between the two pairs of metrics.

Llama3.2-1B. For the Llama model, iterative self-correction generally hinders performance (Fig. 5). In the SQUAD 2.0 dataset, the EM rate, the METEOR score and the BERT F1-score all suffer from iteration 1. These metrics reveal a downward trend with more iterations. In contrast, a positive change is observed

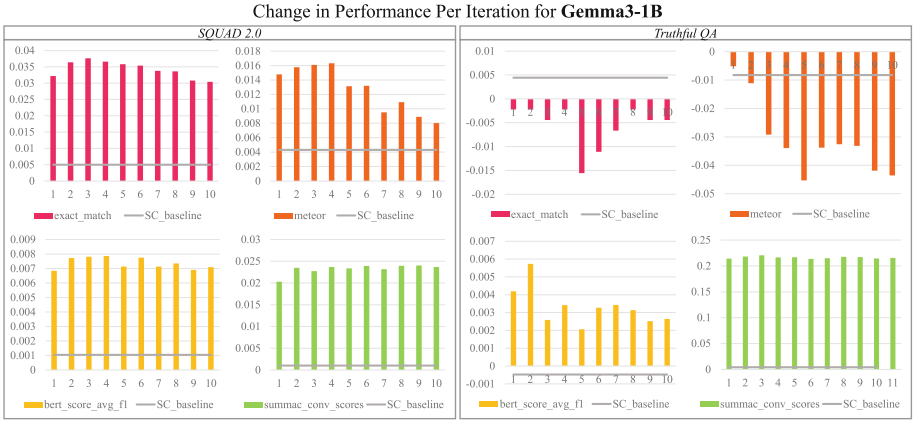


Fig. 4. Changes in the exact match rate (pink), the METEOR score (orange), the BERT F1-score (yellow), and the SummaC score (green) from iteration 0 to each subsequent iteration for the SQUAD 2.0 (left) and TruthfulQA (right) datasets, using Gemma3-1B. (Color figure online)

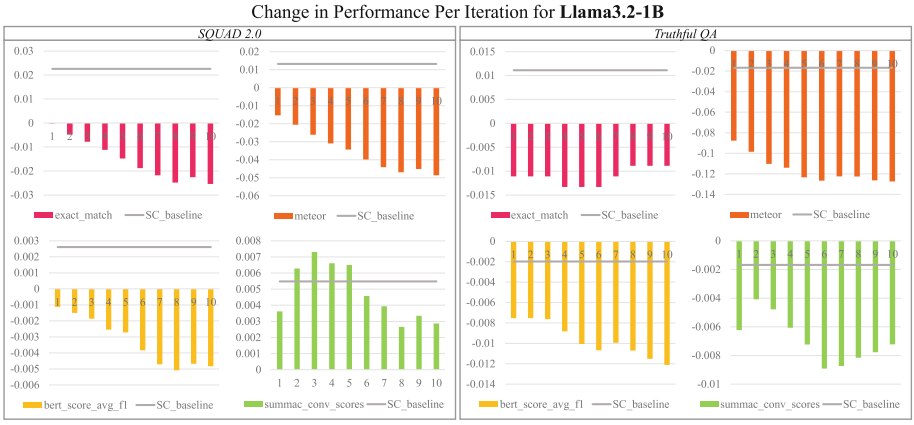


Fig. 5. Changes in the exact match rate (pink), the METEOR score (orange), the BERT F1-score (yellow), and the SummaC score (green) from iteration 0 to each subsequent iteration for the SQUAD 2.0 (left) and TruthfulQA (right) datasets, using Llama3.2-1B. (Color figure online)

for all iterations for the SummaC score, achieving a maximum change by 0.73% at iteration 3 (which exceeds the self-consistency baseline). This observation suggests that the generated responses are likely more abstractive (not lexically or semantically similar to the ground truth), but sheds doubt on whether they are supported by the ground truth due to the low variance of the SummaC score. For the TruthfulQA dataset, a similar trend is observed with the EM rate, the METEOR score, and the BERT F1-score - all displaying a decrease

in performance from iteration 0 for all 10 iterations. The SummaC score also follows suit, displaying a decrease in performance for all 10 iterations and hence confirms the increase of hallucinations after iterative self-correction.

6 Discussion

Based on the experiments conducted, the implementation of iterative self-correction does not entirely work as expected. Although more responses are being retained with every iteration, the ones being kept are both those that were given a perfect score in the feedback and those that were not; the former behaviour is expected in iterative self-correction, while the latter is not. Either the prompt does not advocate for the change in response enough with the quantitative feedback, allowing the model to ignore the feedback altogether, or the descriptive and quantitative versions of the feedback contradict, confusing the model in whether a change in response is required. This is similar to what is observed in [13], where the refinement step using the Vicuna model struggles to adhere to the feedback, returning the same output most of the time. Even though the iterative self-correction mechanism is not behaving completely as expected, the implementation still manages to mitigate hallucinations for the base model. For both the SQUAD 2.0 and TruthfulQA datasets, improvements in scores are observed, suggesting that less hallucinations are being generated with every iteration. Manual analysis of a small subset of examples reveals that the quantitative feedback scores are being incorrectly generated by the model. The descriptive portion of the feedback identifies and reflects on the given response; however it does not match the quantitative feedback being produced. If the previous answer based on the descriptive feedback is perfect (there is no identified flaw in the response), then the quantitative feedback does not assign the traits full scores. As a result, the model in the refinement step does not modify the previous response, favouring the descriptive feedback over the quantitative one. This suggests that the evaluation method that checks the effectiveness of the approach is incomplete - the assumption that both quantitative and descriptive feedbacks match, does not hold. The observed effect on hallucinations reveal that the implementation does, in fact, work for the base model, which, in turn, contradicts the finding from [13] that the strategy does not work for models with less than 13B parameters.

Varying the model size of the Qwen2.5 model achieves contradictory results on hallucinations depending on the dataset being used. The smaller model manages to mitigate hallucinations on the SQUAD 2.0 dataset but struggles for the TruthfulQA one, suggesting that iterative self-correction works better on single-word responses as opposed to more verbose explanations. However, the larger Qwen model directly contradicts that hypothesis; its results show that iterative self-correction mitigates hallucinations better on the TruthfulQA dataset but hinders performance on the SQUAD 2.0 dataset. This discrepancy suggests that A) the capacity of the smaller model is too limited in understanding or producing more complex responses, and B) that the larger model struggles with

providing good quality feedback for situations where the responses are "Cannot be determined". The latter point is suggested from further manual analysis of the results, where both the descriptive and quantitative feedback agree that "Cannot be determined" answers are not relevant to the question, nor consistent with the context (even though they sometimes are), without providing an explicit alternative answer or further explanation. We hypothesize that the larger model may be suffering from the lack of actionable feedback [13] in the SQUAD 2.0 dataset, where the model is unsure how to fix the answer without specific guidance from the feedback, resulting in hallucinations.

Based on the results acquired, the same implementation of iterative self-correction cannot be blindly used on models of different families to achieve the same effect on hallucinations. While the results on Gemma3-1B do show similar improvements throughout all iterations on both datasets, the Llama3.2-1B model does not. We hypothesize that due to the underlying architecture and pre-training, the Llama struggles more at reflecting upon previous answers. Specifically, manual investigation suggests that the Llama model in particular is not following the instructions set by the prompt or properly "understanding" the traits in the feedback, producing feedback that is either not associated with the previous answer, or internally modifying the question so that the answer could be found in the context. Further analysis is required to confirm if changing the prompt or using larger models from these families will have any effect on their generation of hallucinations.

6.1 Limitations and Future Work

This study has two major limitations that may hinder the reliability of results obtained. One is the use of quantized models in the experiments. Quantization allows the models to be used in environments with less computational load, thus allowing the experiments in this study to check larger volumes of data and hence obtain more reliable results. However, these models do contain a trade-off with general performance as, typically, the values stored by the models are limited in magnitude, resulting in some loss of information. Preliminary analysis of the effect of quantization in the Qwen3-0.5B and Llama3.2-1B models reveals no significant difference in the iterative self-correction experiments, achieving results within 1% of the ones reported in this paper. However, to gain a complete understanding of any side effects from quantization, further research in this area is recommended. The second limitation is attributed to the TruthfulQA dataset. Although it allows this study to assess the reasoning capabilities of the model, it is limited in size and the truncation strategy used to extract only the relevant portion of the context may have skewed the models ability in determining the answer; there is no guarantee that the retrieved context actually includes the answer to the question being asked. As such, attempting this experiment on other datasets may shed light onto the generalizability of the iterative self-correction approach.

As part of future work, attempting various prompting strategies such as ones that inspire more actionable feedback, ones that do not use quantitative feed-

back, or ones that use different templates altogether, may reveal the root cause of the unexpected behavior in iterative self-correction and provide a path toward creating more robust and effective implementations. Additionally, our study finds that different models require different number of iterations to obtain the most optimal performance on a dataset. Therefore, further work on the factors that affect how this optimal number varies, may reveal a way to constraint the number of iterations and, thus, allow for more cost-effective implementations of this approach. Furthermore, incorporating an explainability aspect may provide insight into why models of different families behave differently when reflecting on their own response, enabling the development of more tailored approaches that may be better suited for real-world contexts.

7 Conclusion

To sum up, this study confirms the effectiveness of the proposed implementation of iterative self-correction using the base Qwen2.5-1.5B model in an extractive QA setting, revealing a tendency to favour descriptive feedback over quantitative. The strategy decreases the rate of change in the responses with more iterations, allowing answers to be iteratively improved and for hallucinations to be mitigated. The investigation into the model size, reveals that the strategy is effective at improving the performance of smaller models on single-word responses and for larger models on more complex explanations. Any improvements observed by the iterative self-correction strategy exceed that of the self-consistency baseline, shedding light on the efficacy of meta-cognitive strategies in LLMs. Further, the study finds a stark contrast between the effect of iterative self-correction in the Qwen family and that of the Llama family, suggesting that approaches implementing this strategy are tied to the architectural and behavioral characteristics of the model being used. This, in turn, emphasizes the need for tailored approaches and a deeper understanding of how different models generate their own feedback before employing iterative self-correction into real-world contexts.

Disclosure of Interests. The authors of this study have no competing interests to declare that are relevant to the content of this paper.

References

1. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, pp. 65–72. Association for Computational Linguistics (2005). https://www.aclweb.org/anthology/W05-0909
2. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

3. Gou, Z., et al.: Critic: large language models can self-correct with tool-interactive critiquing (2024). https://arxiv.org/abs/2305.11738
4. Huang, J., et al.: Large language models cannot self-correct reasoning yet (2024). https://arxiv.org/abs/2310.01798
5. Huang, L., et al.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**(2) (2025). https://doi.org/10.1145/3703155
6. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023). https://doi.org/10.1145/3571730
7. Kamoi, R., Zhang, Y., Zhang, N., Han, J., Zhang, R.: When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. Trans. Assoc. Comput. Linguist. **12**, 1417–1440 (2024). https://doi.org/10.1162/tacl_a_00713
8. Laban, P., Schnabel, T., Bennett, P.N., Hearst, M.A.: SummaC: re-visiting NLI-based models for inconsistency detection in summarization. Trans. Assoc. Comput. Linguist. **10**, 163–177 (2022). https://doi.org/10.1162/tacl_a_00453
9. Li, Y., Yang, C., Ettinger, A.: When hindsight is not 20/20: testing limits on reflective thinking in large language models. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, pp. 3741–3753. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-naacl.237
10. Lin, J., et al.: AWQ: activation-aware weight quantization for LLM compression and acceleration (2024). https://arxiv.org/abs/2306.00978
11. Lin, S., Hilton, J., Evans, O.: TruthfulQA: measuring how models mimic human falsehoods. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 3214–3252. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.acl-long.229
12. Liu, D., et al.: Large language models have intrinsic self-correction ability (2024). https://arxiv.org/abs/2406.15673
13. Madaan, A., et al.: Self-refine: iterative refinement with self-feedback (2023). https://arxiv.org/abs/2303.17651
14. Manakul, P., Liusie, A., Gales, M.: SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 9004–9017. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.emnlp-main.557
15. Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.173
16. Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., Wang, W.Y.: Automatically correcting large language models: surveying the landscape of diverse automated correction strategies. Trans. Assoc. Comput. Linguist. **12**, 484–506 (2024). https://doi.org/10.1162/tacl_a_00660
17. Portkey, I.: Enhancing truthfulQA with context (2024). https://huggingface.co/datasets/portkey/truthful_qa_context

18. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: unanswerable questions for SQuAD. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, pp. 784–789. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/P18-2124
19. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, pp. 2383–2392. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/D16-1264
20. Renze, M., Guven, E.: Self-reflection in LLM agents: effects on problem-solving performance (2024). https://arxiv.org/abs/2405.06682
21. Sheokand, T., Jain, G., Bahga, A., Madisetti, V.K.: Enhancing LLM reasoning capabilities through brokered multi-expert reflection. IEEE Access **13**, 67993–68019 (2025). https://doi.org/10.1109/ACCESS.2025.3561693
22. Shi, X., Liu, J., Liu, Y., Cheng, Q., Lu, W.: Know where to go: make LLM a relevant, responsible, and trustworthy searchers. Decision Support Syst. **188**, 114354 (2025). https://doi.org/10.1016/j.dss.2024.114354
23. Gemma Team: Gemma 3 (2025). https://goo.gle/Gemma3Report
24. Qwen Team: Qwen2.5: A party of foundation models (2024). https://qwenlm.github.io/blog/qwen2.5/
25. Touvron, H., et al.: Llama: open and efficient foundation language models (2023). https://arxiv.org/abs/2302.13971
26. Yang, A., et al.: Qwen2 technical report (2024). https://arxiv.org/abs/2407.10671
27. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT (2020). https://arxiv.org/abs/1904.09675

Clustering Spatiotemporal Patterns of Infrastructure Investment in Ecuador: A Cantonal-Level Analysis (2008–2022)

Helen Magdalena Gómez Medina[1], Myriam Urbina[2], María Elena Jerez Calero[2], Víctor Cuenca[1], Arcenio Córdova[1], and Erick Cuenca[3](✉)

[1] Departamento de Ciencias Económicas, Administrativas y de Comercio, Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui y Ambato, Sangolquí, Ecuador
{hmgomez,vecuenca,afcordova}@espe.edu.ec

[2] Departamento de Ciencias Económicas, Administrativas y de Comercio Sede Latacunga, Universidad de las Fuerzas Armadas ESPE, Av. Quijano y Ordoñes y Hermanas Paez, Latacunga, Ecuador
{maurbina1,mejerez}@espe.edu.ec

[3] Yachay Tech University, Urcuquí, Ecuador
ecuenca@yachaytech.edu.ec

Abstract. A canton-level geo-spatial analysis of public infrastructure expenditure in Ecuador from 2008 to 2022 is presented. Fiscal records, GIS data, and spatial econometric models are utilized to map the magnitude and geographic distribution of investments and to assess spillover effects among neighboring cantons. Temporal spending patterns are clustered via Dynamic Time Warping and K-Means, resulting in four profiles ranging from negligible disbursements to isolated peaks. Budgeted versus executed funds are subsequently grouped into four performance typologies, highlighting regional disparities in execution efficiency. Local Indicators of Spatial Association (LISA) are applied to classify cantons into "High-High," "High-Low," "Low-High," and "Low-Low" clusters, revealing concentrated zones of high investment and persistent gaps in rural and Amazonian areas. Sectoral analysis indicates that public works and capital transfers predominate. Substantial heterogeneity at the canton level is underscored, and the vulnerability of local finances to external shocks—such as declines in oil revenues and the COVID-19 pandemic—is emphasized. Differentiated policy measures aimed at bolstering budget execution and promoting territorial equity are deemed imperative.

Keywords: Spatiotemporal Distribution · Public Spending · Infrastructure Expenditure

1 Introduction

Infrastructure investment has long been recognised as a catalyst for economic growth and territorial cohesion [13]. Meta-analyses report output multipliers

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 403–418, 2026.
https://doi.org/10.1007/978-3-032-20322-9_28

exceeding unity when projects are efficiently executed, particularly in contexts of acute infrastructure deficits. In Ecuador, fiscal decentralisation has transferred significant capital-budget authority to canton governments; however, outcomes remain uneven. Evidence of β-convergence has been detected only under explicitly redistributive schemes, while marked heterogeneity endures across jurisdictions.

Budget allocations have followed regressive patterns, consistent with the so-called "Matthew effect" noted in the context of university funding [2]. Concurrently, the cantonal share of investment in non-financial assets declined between 2013 and 2023. At the household level, increases in public infrastructure spending are associated with reductions in multidimensional poverty, particularly in rural areas [12].

Micro-level studies further indicate that infrastructure capital explains one-fifth of urban land values and generates owner windfalls multiples higher than initial outlays [7]. Collectively, these findings suggest that both the scale and spatial allocation of infrastructure investment are decisive for equitable development.

Against this backdrop, the present study delivers the first longitudinal, canton-level investigation of Ecuador's infrastructure spending over 2008–2022. The objectives are twofold: (i) to map the magnitude and geography of capital outlays, and (ii) to identify local spatial association patterns among neighbouring cantons using LISA.

2 State of the Art

To situate this study within the broader scholarly discourse, this section synthesizes the state of the art on public infrastructure investment and its socioeconomic impacts. It begins by outlining the conceptual foundations that explicate the channels—productivity spillovers, market integration, and spatial convergence—through which infrastructure shapes regional development. It then presents a PRISMA-based systematic literature review, detailing the search strategy, study selection process, and inclusion/exclusion criteria used to identify empirical analyses in Latin America and Ecuador. Finally, by integrating thematic insights from quantitative spatial-econometric and optimization studies, this overview highlights key methodological advances and enduring gaps that motivate the present research.

2.1 Conceptual Foundations

Three principal channels link public infrastructure to socioeconomic outcomes: (i) productivity spillovers that reduce private production costs; (ii) market integration through lower transaction and transport frictions; and (iii) spatial convergence as networks facilitate resource flows between regions. These mechanisms have been formalized in endogenous growth models and supported by empirical multipliers in digital, energy, and transport sectors.

2.2 PRISMA Systematic Literature Review

Search Strategy. A systematic search was performed in Scopus using a Boolean query combining infrastructure terms (e.g., "public," "infrast*"), financial terms ("invest*," "spend*," "budget*," "expend*"), and Latin American country names, including Ecuador.

1. **Scopus Query**:

```
TITLE-ABS-KEY (
    (public OR infrast*) AND
    ( invest* OR spend* OR budget* OR expend* ) AND
    ( Brazil OR Mexico OR Colombia OR Peru OR Chile OR
        Argentina OR Ecuador OR Bolivia OR Paraguay OR
        Uruguay OR Venezuela ) AND
    ( spatial OR distribution ) )
```

No records fulfilling the eligibility criteria were returned from Dimensions.ai[1], presumably owing to indexing scope differences.

Study Selection: PRISMA Flow Diagram. Figure 1 shows the PRISMA 2020 flow diagram. Of 2,523 Scopus hits, 2,484 were excluded by automated filters. Fifteen records were removed for omitting Ecuador, eight for lacking infrastructure-spending relevance, and six following manual relevance screening. Ten empirical studies satisfied all inclusion criteria.

Inclusion and Exclusion Criteria. A set of eligibility parameters was defined to ensure methodological coherence and temporal comparability across the literature surveyed. Studies were retained when they examined infrastructure spending in Ecuador—or, for contextual breadth, other Latin-American cases—within the 2008–2022 window. Only empirical investigations employing quantitative evidence, preferably spatial-econometric techniques, were considered. Purely theoretical essays, opinion pieces, or case descriptions lacking data were excluded. Finally, the search was restricted to English- and Spanish-language publications, reflecting the dominant scholarly output in the region. The resulting criteria are summarized in Table 1.

Data Extraction and Synthesis. For each study meeting the eligibility criteria, bibliographic metadata were systematically recorded. The geographic scope, methodological framework, and data sources were cataloged. Key quantitative results, policy recommendations, and author-reported caveats were documented. Study quality was appraised by noting sample sizes, temporal coverage, and the

[1] Available at: https://www.dimensions.ai.

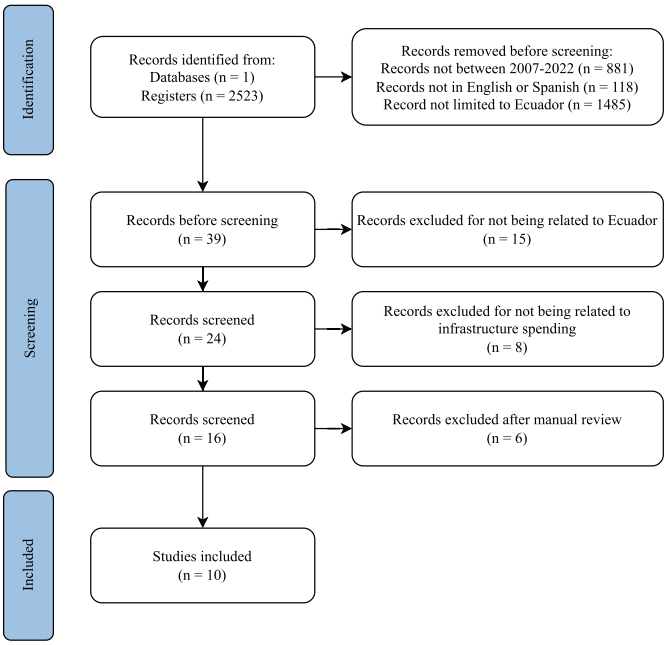


Fig. 1. PRISMA 2020 flow diagram of study selection process.

Table 1. Eligibility criteria applied during study selection

Criterion	Inclusion	Exclusion
Geography	Ecuador; Latin America for expanded context	Non–Latin American contexts
Time Frame	2008–2022	Published before 2008 or after 2022
Study Type	Empirical analyses; spatial econometrics	Purely theoretical; opinion pieces; case reports without data
Language	English or Spanish	Other languages

use of robustness checks or sensitivity analyses. Subsequently, thematic grouping was performed to isolate evidence of spatial-dependence effects in budget execution and socioeconomic outcomes, alongside the deployment of optimization and multi-criteria decision-making techniques for resource allocation.

Summary of Findings. Table 2 distills each study's principal contributions and methodological or contextual constraints. In the "Key Findings" column, how spatial-econometric tools uncovered regional spillovers, convergence patterns, or optimization trade-offs is highlighted. In the "Limitations" column, recurring issues—such as reliance on cross-sectional designs, scale restrictions, absence of causal identification strategies, and case-study narrowness—are flagged.

Table 2. Summary of Findings from the 10 Included Studies

Reference	Key Findings	Limitations
[18]	Quality-management tools applied to thematic geoinformation improved the precision and efficiency of municipal planning in various Ecuadorian cantons.	Analysis limited to scale 1:25,000; generalizability to other scales or regions not assessed.
[10]	Urbanization positively correlates with per-capita regional output in Ecuador; spatial lag and error models reveal significant spillover effects.	Provincial-level data only; potential endogeneity despite instrumental-variable correction.
[15]	A multi-criteria decision-making framework enhanced productivity metrics in Cuenca's public-transport network by prioritizing service upgrades.	Case study limited to Cuenca; transferability to other cities not empirically tested.
[11]	Genetic-algorithm optimization balanced cost, equity, and reliability in intermittent water-supply investments for Milagro canton.	Model assumptions (e.g. weightings, demand patterns) may not hold in other supply contexts.
[7]	Public infrastructure investment in Cuenca drove significant increases in urban land prices; elasticities were estimated via hedonic models.	Single-city focus; omitted variable bias possible in hedonic specifications.
[12]	Higher public spending categories (health, education) are associated with lower multidimensional poverty indices across Ecuadorian cantons.	Aggregate annual data; causality cannot be firmly established.
[17]	An integrated, systemic planning approach supported sustainable tourism development in Bahía de Caráquez by aligning infrastructure, environmental, and social objectives.	Predominantly qualitative; quantitative performance metrics were not rigorously evaluated.
[14]	Parametric and nonparametric models provide evidence of both conditional and unconditional regional economic convergence in Ecuador.	Focused on income convergence; direct links to infrastructure spending not explicitly modeled.
[3]	Technification of irrigation in the Pisque River Basin improved agricultural productivity and community resilience.	Narrow environmental scope; long-term sustainability impacts not assessed.
[6]	Spatial-econometric analysis shows that public expenditure contributes to reducing regional disparities in Ecuador via spillover effects.	Data aggregated at the provincial level; canton-level heterogeneity remains unexplored.

The ten studies collectively demonstrate that spatial-econometric methods uncover significant inter-regional spillovers in infrastructure outcomes; that multi-objective optimization frameworks effectively balance equity and efficiency in water and transport investments; and that integrating geospatial data with econometric models enhances analytical precision, albeit constrained by heterogeneous data sources and context-specific limitations.

International Perspective (Concise). Beyond Latin America, recent work employs clustering to model infrastructure development dynamics and their spatial context. For example, Kussul et al. use clustering to profile village-level infrastructure trajectories, speaking to the value of interpretable cluster archetypes for planning [8]; and Wu et al. analyze urban infrastructure impacts on urbanization from a spatiotemporal perspective, reinforcing the utility of combining spatial structure with temporal profiles [16]. These studies align with our modular approach (spatial clusters and LISA; temporal DTW profiles) while keeping policy interpretability front and center.

2.3 Previous Studies in Latin America and Ecuador

Latin America. Several multi-country analyses have underscored the pivotal role of infrastructure in Latin American growth. Easterly, Calderón, and Servén [13] found that road-network density and electrical generation capacity explain a substantial share of cross-country GDP differentials, highlighting the importance of accumulated infrastructure assets for long-term convergence. The Inter-American Development Bank [4] estimated that the region will need annual investment equivalent to 5–6% of GDP through 2030 to meet Sustainable Development Goals, emphasizing both financing mobilization and efficiency improvements in public spending.

Ecuador. At the national level, research on Ecuador's public spending has employed both econometric and optimization approaches. Álvarez-García et al. [1] constructed a Global Productive Infrastructure Index (GPII) to show that transport and communications infrastructure are the strongest predictors of provincial growth rates. A UNACH study analyzed road-infrastructure expenditure from 2010 to 2021, finding significant positive effects on regional GDP, but also revealing persistent execution gaps due to administrative bottlenecks [9]. More sector-specific work includes:

- **Urban Mobility**: Cazorla and Ávila Ordoñez [15] assessed baseline conditions in Cuenca's public-transit network and proposed prioritization schemes to improve service coverage and reliability.
- **Geospatial Planning**: Zapata et al. [18] applied quality-management tools to geoinformation in municipal planning, demonstrating that standardized spatial data increases the accuracy of budget-allocation models.

– **Water-Supply Optimization**: Nyahora et al. [11] developed a multi-objective genetic-algorithm framework to balance cost, equity, and reliability in intermittent water-supply investments for Milagro canton.

2.4 Research Gap Addressed by This Study

Despite this body of work, few analyses investigate infrastructure-spending dynamics at the canton level in Ecuador over multiple budget cycles. Existing studies often use cross-sectional snapshots or focus narrowly on single sectors, limiting insights into spatial spillovers and temporal evolution [5]. Moreover, data heterogeneity and inconsistent reporting standards across ministries impede comprehensive, longitudinal assessments. This study bridges these gaps by harmonizing administrative records across ministries to assemble a canton-level annual panel (2008–2022), enabling joint spatial and temporal analyses that can inform targeted policy interventions and equitable resource allocation.

3 Methodology

This study implements a **two-module** analysis pipeline. First, a *spatial* module combines K-Means clustering on neighbor-weighted panel means with local spatial autocorrelation (LISA) to summarize geographic heterogeneity and dependence. Second, a *temporal* module profiles canton-level execution trajectories via DTW + K-Means. The sections that follow mirror this organization in the Results: *Spatial Clustering and Autocorrelation*, and *Temporal Clustering*.

3.1 Data Sources and Preparation

We assemble a balanced panel for **224 cantons** over **2008–2022** from administrative budget and expenditure records (Ministry of Finance; canton-level, not publicly released) and official canton boundaries (GeoJSON; WGS 84).[2] A Queen contiguity structure is built from the canton geometries for spatial analysis.

Ingestion and Cleaning. Raw tables were exported to CSV and read with `pandas`. Columns were standardized to lowercase with underscores, harmonizing synonyms (e.g., *budget_ amt* / *allocated* → `budgeted`). Types were enforced (`year=int64`, monetary fields `float64`). After basic consistency checks, we constructed canton × year aggregates for budgeted and executed amounts and merged them to the GeoJSON by standardized canton identifiers under WGS 84. No additional outlier trimming or winsorization was applied.

EDA Overview. Table 3 summarizes the structure. The analytic table contains **163,436 rows × 12 columns**; required identifiers (group, item, province, canton) are fully populated and monetary fields are nonnegative.

[2] GeoJSON source: Emamut (2018), https://gist.github.com/emamut/25912e117ab46fa00a63c6e890575201.

Table 3. EDA overview of dataset structure (shape: 163,436 rows × 12 columns).

Column	Data Type	Missing Values	Unique Values
year	int64	0	15
group_id	int64	0	10
group_desc	str	0	8
item_id	int64	0	332
item_desc	str	0	669
province_code	str	0	40
province	str	0	41
canton_code	str	0	282
canton	str	0	267
udaf	str	0	371
budgeted	float64	0	–
spent	float64	0	–

Aggregate Definitions. Let i index cantons and t years. Budgeted and executed totals for canton–year are

$$\text{TotalBudgeted}_{i,t} = \sum_{r \in (i,t)} \text{budgeted}_r, \qquad \text{TotalSpent}_{i,t} = \sum_{r \in (i,t)} \text{spent}_r.$$

These aggregates are used for mapping, spatial statistics, and time-series profiling after the spatial join to the canton geometries.

3.2 Analytical Strategy

We adopt a **modular** pipeline with two independent yet complementary components: a *spatial* module (K-Means + LISA) to summarize geography and local spatial association patterns, and a *temporal* module (DTW + K-Means) to profile yearly execution trajectories. This separation preserves interpretability for policy audiences while allowing each module to use the most appropriate distance and inference tools.

Spatial Clustering (K-Means). We partition cantons with **K-Means** on neighbor-weighted panel means of budget and spending (2008–2022). The number of clusters k is selected by the Calinski–Harabasz index. Implementation settings: `StandardScaler()`, `KMeans(n_clusters=4, random_state=0)`. Resulting labels feed the map in Fig. 2a.

Local Spatial Autocorrelation (LISA). We compute LISA on the **panel-average executed spending** (TotalSpent_i, 2008–2022) with **999 permutations** and a **5%** pseudo-p threshold, applying Benjamini–Hochberg FDR correction across cantons. Global Moran's I on the same variable summarizes overall spatial dependence.

Temporal Clustering (DTW + K-Means). Annual execution series are standardized (z-scores). We cluster with **TimeSeriesKMeans** using the **DTW** metric (`random_state=0`). The number of clusters k is chosen via the elbow method corroborated by average silhouette, and we summarize each group by its *DTW centroid*.

Why DTW Despite Equal Series Length? Execution peaks often occur in *different calendar years* across cantons due to procurement lags, emergency reprioritizations, and exogenous shocks. DTW aligns such *phase shifts*, emphasizing *shape* rather than *amplitude*.

4 Results and Discussion

This section presents the key empirical findings and their interpretation, beginning with an exploration of spatial patterns in budget allocation and execution across cantons, followed by an analysis of temporal dynamics in national infrastructure spending over the 2008–2022 period. We then examine the distribution of investment across infrastructure categories to identify dominant sectors and conclude by integrating these insights to discuss the implications of observed heterogeneity, spillover effects, and policy-relevant intervention points.

4.1 Spatial Clustering and Autocorrelation

A neighbor-weighted comparison of mean budget allocations versus actual spending over the 2008–2022 period reveals a broad "middle band" of cantons, where average budgets of USD 8–12 million correspond to execution rates close to 90%. However, substantial heterogeneity is also apparent. A subset of cantons persistently under-spends relative to their allotments, suggesting administrative bottlenecks or limited project pipelines, while another group secures disproportionately large inflows—often linked to strategic infrastructure programs or external grants—relative to neighboring jurisdictions. To formalize these patterns, cantons were clustered according to their neighbor-weighted mean budget and execution values, yielding four distinct spatial profiles:

- **Budget and moderate spending:** These units exhibit stable execution rates around 90% on budgets ranging between USD 8 million and USD 12 million. They are typically found in mid-sized urban centers and secondary provincial capitals, where fiscal management practices and project pipelines allow for predictable disbursement of funds.
- **High expense and budget:** Concentrated in primary metropolitan hubs such as Quito and Guayaquil, these cantons consistently register both budget and spending levels above USD 20 million. Their elevated capacity for large-scale infrastructure works reflects stronger institutional frameworks, higher population densities, and greater access to diversified revenue streams.

- **Execution vs. budget mismatch:** Located primarily along the southern coast and in isolated highland valleys, these cantons display allocations that exceed actual expenditures by 30–50%. The persistent gap points to logistical challenges, protracted procurement processes, or limited local implementation capacity that hinder full absorption of authorized funds.
- **High spending/budget outliers:** This outlier canton records an execution-to-budget ratio in excess of 1.5, driven by emergency relief projects or large donor-funded initiatives that inflate spending beyond planned budgets. Such exceptional cases underscore the influence of ad hoc interventions and highlight the need to incorporate non–budgetary funding streams into routine oversight.

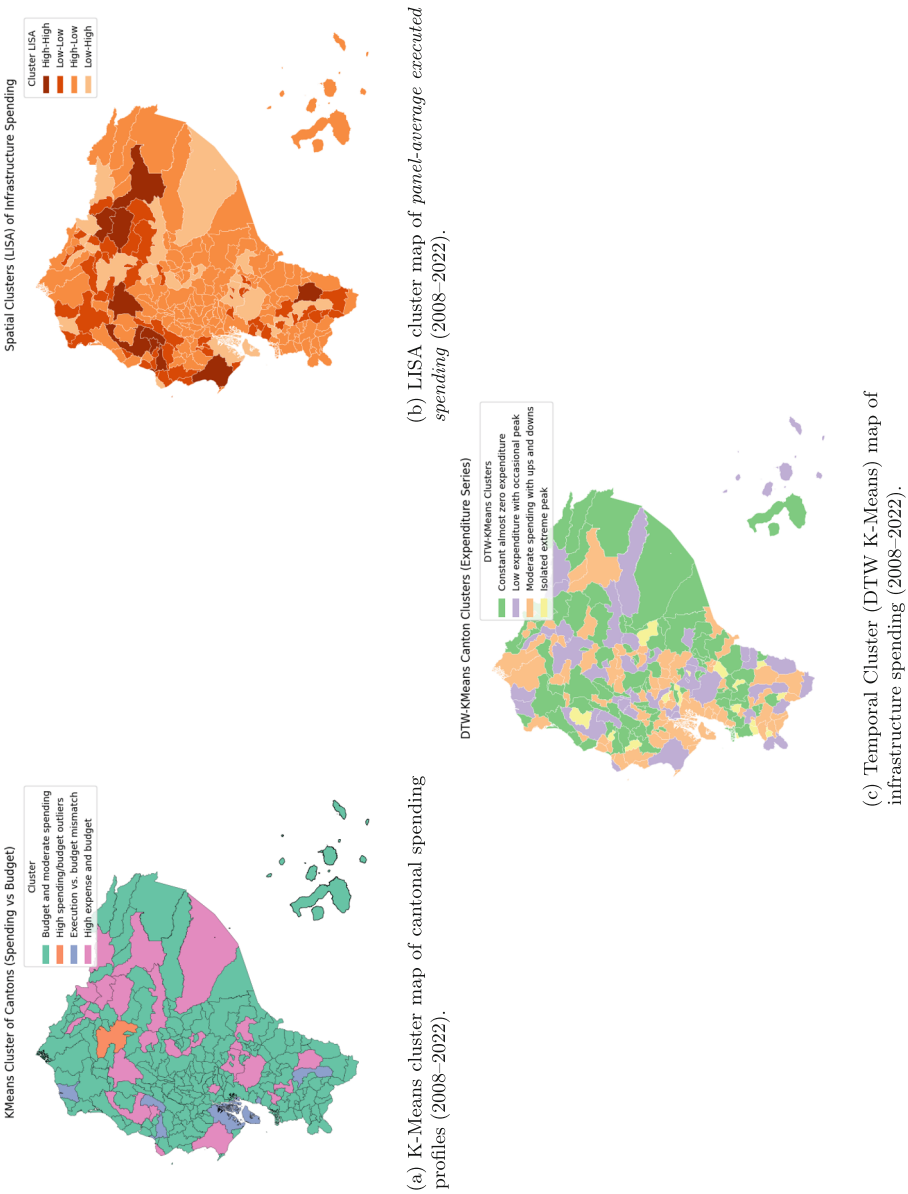
The K-Means Cluster of Cantons map (Fig. 2a) complements the temporal clustering by highlighting regions that exhibit discrepancies in budget implementation. These mismatches suggest a need for canton-specific policies—tailored to local realities—that facilitate the execution of targeted projects with tangible socioeconomic impact.

The Spatial Cluster (LISA) of Infrastructure Spending map (Fig. 2b) reveals the existence of economic development corridors or zones where high infrastructure investment coincides with cantons characterized by high population density and key transportation or regional-planning nodes. Such clusters may generate spillover benefits for neighboring cantons. The Low-High cluster, in particular, signals focal points of investment; most adjacent cantons fail to replicate this level of spending due to factors such as inequitable resource distribution and structural technical or administrative barriers, thereby underscoring the necessity of differentiated territorial intervention policies.

- **High-High:** Clusters around Quito, Manabí, Guayas, Orellana, Napo—both canton and neighbors post above-average spending.
- **High-Low:** Cantons (e.g., Galápagos) with high spending amid low-spending neighbors.
- **Low-High:** Under-invested rural cantons surrounded by high-investment neighbors.
- **Low-Low:** Amazon & remote highlands where both canton and neighbors persistently under-spend
- **Not significant:** cantons without statistically significant local association at the 5% FDR-adjusted level.

4.2 Temporal Clustering (DTW + K-Means)

Between 2008 and 2012, the authorized infrastructure budget for Ecuador increased from approximately USD 2 billion to nearly USD 4 billion, a trend driven by elevated oil revenues and an expansionary fiscal stance. During this interval, actual expenditures closely mirrored approved allocations, yielding mean execution rates of 87.5% (standard deviation: 3.2%), as illustrated in Fig. 3.



(a) K-Means cluster map of cantonal spending profiles (2008–2022).

(b) LISA cluster map of *panel-average executed spending* (2008–2022).

(c) Temporal Cluster (DTW K-Means) map of infrastructure spending (2008–2022).

Fig. 2. Comparison of spatial K-Means, LISA clusters, and temporal DTW clusters.

Line-and-boxplot visualizations reveal that the inter-annual variation in both budgeted and executed amounts remained comparatively modest, with interquartile ranges of USD 0.3 billion for budgets and USD 0.25 billion for spending.

Following a zenith in 2013–2014—when budget authorizations peaked at roughly USD 4.2 billion and outlays reached USD 3.8 billion—a sustained contraction was observed. From 2015 to 2019, a cumulative reduction of 15% in both budget and spending was recorded, correlating with the global downturn in oil prices and the implementation of austerity measures. Execution rates during this period oscillated between 70% and 95%, with a coefficient of variation of 12%, reflecting intermittent fiscal tightening and loosening cycles.

The advent of the COVID-19 pandemic in 2020 precipitated a sharp decline: execution of infrastructure budgets fell by 25% year-on-year despite a modest increase in emergency appropriations. This interruption was followed by a 2021 stimulus package that temporarily restored execution rates to 95% of allocated funds, indicating rapid mobilization of recovery-oriented spending. By 2022, however, both budgeted and actual expenditures stabilized at levels approximately 30% below the 2014 peaks, underscoring the persistent vulnerability of infrastructure financing to external shocks.

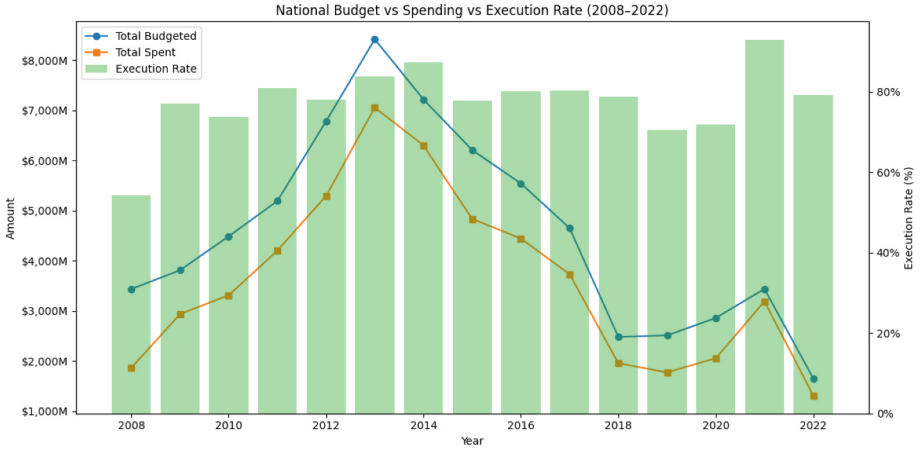


Fig. 3. National infrastructure spending: **budget (left axis: USD billions)**, spending (left axis: USD billions), and execution rate (right axis: %) for 2008–2022.

To uncover canton-level heterogeneity in these national trends, Dynamic Time Warping combined with K-Means clustering was applied to each canton's normalized execution time series. Four distinct temporal profiles emerged (see Fig. 2c):

– **Constant almost zero expenditure:** These units, predominantly located in the Amazon and Galápagos regions, maintained annual outlays below USD 1 million, with minimal temporal fluctuation (variance < 0.1).
– **Isolated extreme peak:** Characterized by singular spikes—often tied to emergency or donor-funded works—followed by baseline spending below USD

2 million; peak magnitudes exceeded three standard deviations above the mean.
- **Low expenditure with occasional peaks:** Exhibited base spending of USD 3–5 million, punctuated by medium-scale project launches yielding spikes between USD 10 million and USD 15 million; silhouette widths averaged 0.62, indicating clear cluster separation.
- **Moderate spending with ups and downs:** Comprised mostly urban and peri-urban areas such as Guayaquil and Quito's suburbs, maintaining annual expenditures of USD 10–20 million and regular oscillations with a period of 2–3 years.

When the DTW–KMeans clustering results are mapped onto the cantonal geography, a pronounced spatial correspondence with levels of socioeconomic development becomes evident. Cantons assigned to the "Moderate spending with ups and downs" profile are predominantly located in metropolitan and peri-urban zones—such as the outskirts of Quito and Guayaquil—where robust administrative capacity and larger tax bases support sustained investment flows. In contrast, cantons falling into the "Isolated extreme peak" cluster are scattered across remote highland valleys and Amazonian fringes, where episodic emergency works create one-off spending spikes against a background of otherwise minimal annual allocations. These contrasting spatial distributions suggest that institutional factors, project pipeline management, and accessibility constraints play critical roles in shaping temporal expenditure trajectories. Furthermore, the magnitude and timing of budget disbursements differ substantially between planning and execution phases: more developed cantons exhibit smoother transitions from approval to disbursement, whereas less developed areas often encounter administrative delays or intermittent funding windows. Such heterogeneity underscores the necessity of tailoring fiscal oversight and technical assistance to the unique capacity profiles of each canton.

4.3 Infrastructure Category Distribution

A detailed breakdown of cumulative investment by category over the 2008–2022 period (Fig. 4) reveals a clear predominance of public works expenditures. Within this grouping, transportation infrastructure—comprising road rehabilitation, bridge construction, and port upgrades—accounts for the largest share of outlays, followed by investments in public buildings such as schools, health centers, and municipal administrative facilities. Water supply and sanitation projects represent the next most significant allocation, reflecting longstanding efforts to expand access to potable water and improve wastewater treatment in both urban and rural contexts. Investment transfers and donations emerge as the second most substantial category, encompassing intergovernmental capital grants and externally funded projects that do not directly tie to canton-level procurement processes. These transfers are often used to support sector-specific initiatives—such as renewable energy pilot schemes or digital connectivity programs—without passing through the standard budget execution pipeline.

Lesser shares are directed toward sectors like energy generation, irrigation infrastructure, and social housing, highlighting both the diversification of infrastructure priorities and the influence of external funding sources on cantonal spending patterns.

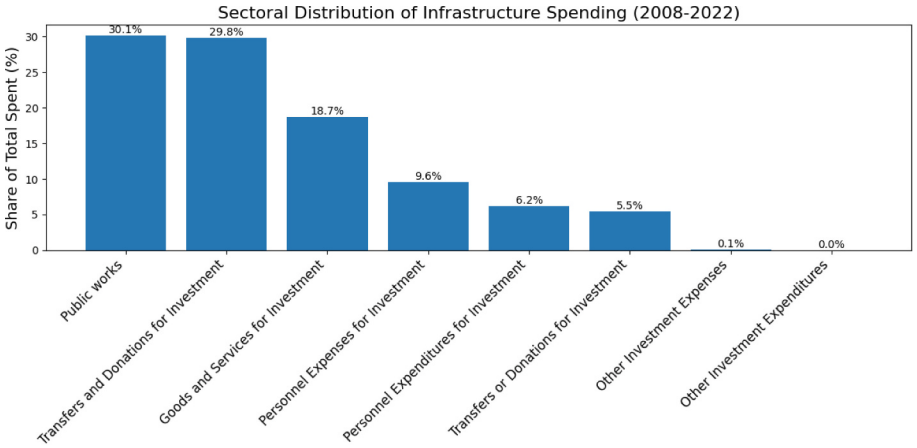


Fig. 4. Sectoral distribution of infrastructure spending in Ecuadorian cantons (2008–2022).

It has been demonstrated by Zapata et al. [18] that spatial standardization enhances accuracy, albeit only at a 1:25000 scale. In contrast, precision has been extended to the cantonal level, and multiple budgetary cycles have been encompassed by the 2008–2022 panel. Furthermore, the relationship observed between urbanization and regional per-capita output at the provincial level is confirmed, and spillover effects within cantons are further detailed, thereby revealing internal heterogeneity that had not been charted before. Transport prioritization in Cuenca was examined by Cazorla and Ávila Ordóñez [15], and it was shown why "Moderate spending" clusters are concentrated in well-connected urban nodes. Similarly, genetic algorithms were applied by Nyahora et al. [11] to balance equity and cost in Milagro; this approach has been generalized nationwide through the incorporation of annual spending dynamics via a temporal DTW–KMeans methodology. Finally, although provincial-level spillovers have been documented in earlier spatial-econometric analyses, a deeper understanding of this relationship has been achieved by employing the longitudinal cantonal dataset, which captures heterogeneity and temporal evolution that had not been previously explored.

5 Conclusions and Recommendations

The budgeted amounts and actual expenditures on infrastructure over the 2008–2022 period exhibit a rising trend up to 2013–2014, followed by a pronounced and sustained decline, particularly in 2018 and 2021, years marked by reduced

appropriations. The highest execution rate coincides with the boom period, yet historical analysis reveals considerable volatility. Macroeconomic events, such as economic crises and fiscal adjustments, impair the state's investment capacity, resulting in lower budget allocations to canton governments.

This analysis combines time-series and spatial techniques: Dynamic Time Warping (DTW) paired with K-Means clustering to identify temporal spending patterns; K-Means on neighbor-weighted mean budget versus execution to uncover spatial profiles; and Local Indicators of Spatial Association (LISA) for hot-spot detection. Sectoral shares were computed from the categorization of annual expenditures, and all figures were generated using these methods.

Overall, the results demonstrate substantial heterogeneity in actual spending among cantons over time, indicating that each canton faces distinct circumstances; consequently, uniform policy regulations adversely affect the overall execution rate. In particular, cantons classified within clusters characterized by persistently minimal spending or low expenditures with occasional peaks require the implementation of targeted support programs and institutional strengthening measures to incentivize investment and ensure spending sustainability—measures that would ultimately yield positive impacts on socioeconomic indicators.

Spatial clustering analysis reveals geographic concentrations in budget management, with a dominant cluster of moderate budgets and expenditures. The LISA analysis confirms significant spatial autocorrelation. There are geographic zones of high investment that generate positive spillover effects; nevertheless, many low-spending cantons surrounded by high-spending neighbors remain unable to match their peers in budget execution.

Limitations and Future Work. We report spatial and temporal clustering *separately* to preserve interpretability; fully joint spatiotemporal models could uncover higher-order interactions at the cost of readability. Our DTW choice privileges *shape* under plausible phase shifts. Finally, sectoral categories follow administered labels; future work may re-classify items (e.g., with NLP) to validate cross-ministerial consistency.

Acknowledgments. This work was supported by the University of the Armed Forces, Ecuador, under the program "2024 Applied Research Project" (Grant No: 2024-PIS-02).

References

1. Alvarez Pagliuca, C., Martínez Álvarez, J., Pereira Dos Santos, P., Serebrisky, T., Suárez-Alemán, A.: Financiamiento sostenible de la infraestructura económica y social en américa latina y el caribe: tendencias, actores e instrumentos (2022)
2. Briones, V.V., Morales, J.C., Vásconez, G., Cruzatty, L.G.: Injusta distribución del presupuesto entre las universidades públicas del ecuador: Efecto mateo. Revista Internacional de Educación para la Justicia Social **10**(1) (2021)

3. Cachipuendo, C.: The technification of irrigation as a strategy of community resilience. case study: Pisque river basin, Ecuador. Smart Innov. Syst. Technol. **252**, 207–216 (2022)
4. de Desarrollo, B.I.: Informe anual del banco interamericano de desarrollo 2019: Reseña del año (2020)
5. Fay, M., Martimort, D., Straub, S.: Funding and financing infrastructure: the joint-use of public and private finance. J. Dev. Econ. **150**, 102629 (2021)
6. Flores-Chamba, J., Correa-Quezada, R., álvarez García, J., del Río-Rama, M.d.l.C.: Spatial economic convergence and public expenditure in Ecuador. Symmetry **11**(2) (2019)
7. Iñeguez, J.B., Auquilla, C.C.: Effect of public infrastructure on urban land price: case of the city of Cuenca, Ecuador. Urbano **24**(43), 74–83 (2021)
8. Kussul, N., Potuzhnyi, B., Svirsh, V.: Clustering techniques for modeling village infrastructure development. In: Proceedings of the International Conference on Computational Linguistics and Intelligent Systems (COLINS), vol. 2, pp. 98–119 (2024)
9. Maria José, R.S.: Infraestructura vial y condiciones socioeconómicas: un análisis de conglomerados de la parroquia de quimiag (2020)
10. Méndez, P., Guarnizo, S., Alvarado, R.: Nexo causal entre la urbanización y la producción regional: evidencia para ecuador usando vi en econometría espacial. Reg. Sect. Econ. Stud. **19**(1), 63–76 (2019)
11. Nyahora, P.P., Babel, M.S., Ferras, D., Emen, A.: Multi-objective optimization for improving equity and reliability in intermittent water supply systems. Water Supply **20**(5), 1592–1603 (2020)
12. Rosales, V.Q., Leverone, M.B., Vargas, M.S., Murillo, C.M.: Multidimensional poverty index and its relationship with Ecuadorian public spending. Universidad y Sociedad **12**(2), 430–436 (2020)
13. Servén, L., Calderón, C.: The effects of infrastructure development on growth and income distribution. Documentos de Trabajo (Banco Central de Chile) **270**, 1–47 (2004)
14. Szeles, M., Mendietamuñoz, R.: Analyzing the regional economic convergence in Ecuador. Insights from parametric and nonparametric models. Rom. J. Econ. Forecast. **19**(2), 43–65 (2016)
15. Vanegas, N.P.C.: Un proceso holístico de toma de decisiones para mejorar la productividad del transporte público en cuenca-ecuador. Revista Politécnica **48**(2), 33–42 (2021)
16. Wu, Y., Qian, P., Yang, L., Tian, Z., Luo, J.: Analysis of the impact of urban infrastructure on urbanization processes at different levels from a spatiotemporal perspective. Sustainability **16**(16) (2024)
17. Zambrano, M., Hechavarría Hernández, J.R.: Systemic approach to strategic tourism planning in the cantonal capital of bahía de Caráquez, Sucre, Ecuador. Adv. Intell. Syst. Comput. **1131 AISC**, 1206 – 1212 (2020)
18. Zapata, A., et al.: Application of quality tools for evaluation of the use of geo-information in various municipalities of Ecuador. In: Rodriguez Morales, G., Fonseca C., E.R., Salgado, J.P., Pérez-Gosende, P., Orellana Cordero, M., Berrezueta, S. (eds.) TICEC 2020. CCIS, vol. 1307, pp. 420–433. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62833-8_31

Where Is the Research Heading? A Text Mining Analysis of International Business Publications in Peru

Hugo Giraldo, Eva Farromeque, Adriana Ticse, Grycel Ticse, Julio Ricardo Moscoso Cuaresma(✉), Miguel Nunez-del-Prado, Juan Eduardo Acosta Mantaro, and Carlos Alberto Azabache Morán

Universidad Peruana de Ciencias Aplicadas, Prolongación Primavera 2390, Santiago de Surco, Lima, Peru
julio.moscoso@upc.pe

Abstract. The discipline of international business has become increasingly important in Peru, driven by globalisation and trade liberalisation. This study aims to provide a comprehensive overview of the main subject areas covered in undergraduate academic work in International Business, identifying both established lines of research and those that require further development. To this end, a bibliometric analysis was carried out of theses defended between 2020 and 2024 to obtain a bachelor's degree at Peruvian universities, using the RENATI managed by SUNEDU as a source. The results show a high concentration of research on exports, international logistics, e-commerce and trade agreements, with a predominance of quantitative and descriptive methodological approaches. However, significant limitations were identified, such as heterogeneity in the quality and structure of the work, limited availability of primary data and the scarce presence of innovative theoretical and methodological approaches. Despite these limitations, RENATI is consolidating its position as a valuable source for analysing national academic output, enabling the identification of trends, thematic gaps and opportunities for strengthening educational research in the field. This assessment applies text mining techniques to contribute to the design of more relevant academic strategies for students, graduates and professors of international business in the face of the contemporary challenges of the global and regional environment currently demanded by academic literature.

Keywords: International Business · Thesis Production · Peruvian Universities · Bibliometric Analysis · Text Mining · Co-occurence graphs

1 Introduction

The free trade paradigm, understood as the progressive elimination of tariffs and barriers to trade, was consolidated after the Cold War with the impetus of organisations such as the World Trade Organisation (WTO) and the World

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 419–432, 2026.
https://doi.org/10.1007/978-3-032-20322-9_29

Bank [1], which highlighted its benefits in terms of productive efficiency, the attraction of foreign investment and export diversification. In Latin America, various governments adopted this approach to stabilise indebted economies and gain preferential access to developed markets through bilateral or regional agreements [3]. This liberal logic reconfigured public policies and national priorities in competitiveness, infrastructure and human capital formation.

In Peru, these ideas were translated into structural reforms during the 1990s [4]. Between 1991 and 1996, the average tariff rate fell from over 60% to less than 15% [5]. Accession to the WTO in 1995 and entry into APEC in 1998 deepened the opening [6,7]. At the same time, free trade agreements raised the trade-to-GDP ratio from 27% in 1990 to over 45% in the 2000s [8], while foreign direct investment went from negative figures to over USD 2 billion per year [9]. This context forced Peruvian companies to integrate into global chains subject to demanding technical and regulatory standards.

This opening up increased the need for skilled human capital to lead international operations [10], take financial risks and operate in multicultural contexts [11]. Universities responded by expanding and specialising their international business programmes, which went from being minor subjects in administration to independent degrees with courses in international logistics, comparative commercial law, global finance and intercultural marketing, aligned with an export-led growth model.

Since the early 2000s, the degree programme has expanded rapidly. Universities such as the Universidad del Pacífico, PUCP, Lima and UPC led this trend, which was joined by regional institutions. The signing of trade agreements, integration into global value chains and the sustained increase in foreign investment drove this growth. Curricula evolved to incorporate sustainability, social responsibility and digitalisation, in line with global standards [12].

At the same time, research in International Business grew steadily. In 2024, RENATI housed more than 7,000 theses on internationalisation, sustainability, or digital transformation [13], driven by University Law No. 30220, which established research as an essential function and required the open registration of theses [14]. According to SUNEDU [15], Peruvian indexed articles increased from 1,187 in 2013 to 7,361 in 2022, a 520% increase. Peru climbed from 91st to 71st in the Scimago Country Rank, thanks to incentives from CONCYTEC and the ALICIA repository. In social sciences, the average annual growth was 28%, with 37% of publications co-authored internationally [16].

However, obstacles to developing rigorous theses remain. Restricted access to databases, outdated statistics, scarce field resources, and difficulty in finding key informants [17,18] limit innovation. Added to this is the lack of advanced mentoring and tight academic deadlines, which prevent the formulation of original problems and the application of demanding methodologies [19].

Furthermore, there is a thematic gap in Latin American research on international business. Studies on structural development, sustainability, or competitiveness are scarce, which represents both a weakness and an opportunity [20]. This study provides a theoretical and practical basis for future research on inter-

national trade and management and seeks to guide institutional efforts toward greater academic and scientific impact.

2 Methodology

This research was conducted between March and July 2025 and consisted of a data mining-based bibliometric analysis of bachelor's theses in international business presented at Peruvian universities between 2020 and 2024. The main source used was the National Repository of Research Papers (RENATI), a digital platform managed by the National Superintendency of Higher University Education (SUNEDU), which centralises and openly disseminates theses and academic papers produced by university students in Peru. This repository not only seeks to safeguard university intellectual production, but also to promote academic transparency, facilitate access to knowledge, and contribute to improving research quality [13–15].

RENATI was selected as the official national repository, which ensures that the information collected is representative of academic production in the field of International Business. Unlike other platforms such as Google Scholar, ALICIA or institutional repositories, RENATI focuses exclusively on theses produced by Peruvian universities, which allows it to accurately reflect the predominant themes and approaches in this discipline. However, the use of the repository revealed certain limitations related to the accuracy of the thematic classification of the theses. When applying filters by degree programme, cases were detected in which works belonging to other specialities such as Marketing or Human Resources were included. This situation highlighted the need to improve the thesis labelling and registration processes, both by the universities and by the system that administers the platform.

During data collection, the RENATI website (https://renati.sunedu.gob.pe/) was accessed and filters were applied by degree programme, country of origin (Peru) and year of publication, limiting the range to the period 2020–2024. A total of 2,058 theses registered as corresponding to International Business were identified. However, after a cleansing process that excluded misclassified or non-degree-related works (such as reports or preliminary projects), 10 theses were discarded, leaving a valid database of 2,048 documents. This initial screening also included the exclusion of embargoed or restricted-access theses, a common practice in academic settings that temporarily limits access to the full content of the work to protect exclusive findings or lines of research still under development [21].

The valid records were exported in CSV format and organised into an Excel template designed for preliminary analysis. Key variables were extracted, such as the name of the issuing university, institutions cited, title of the work, year of publication, number of authors, advisors involved, keywords, type of academic degree, and disciplinary classification. Prior to analysis, university names were standardised due to the presence of abbreviations, spelling errors, and incorrect translations. Discrepancies between the thematic classification assigned by the

system and the actual content of the thesis were also corrected, ensuring greater consistency and reliability in the analysed data.

The data were processed using two complementary tools: the Python programming language and Microsoft Excel spreadsheets. Python allowed for the automation of data cleaning, word clouds, terms co-occurrence, topic modelling, and visualisation, using libraries such as PyLDA, Pandas, NetworkX, Matplotlib, and Seaborn for the creation of graphs and visual representations. Excel facilitated the manual validation of records and the construction of dynamic tables. Among the aspects analysed, the universities with the highest volume of published theses, the most cited institutions, the advisors with the highest participation, the frequency of recurring authors, the distribution of theses by year, type of degree (bachelor's or master's), and the recurrence of keywords stood out. This last analysis made it possible to identify common thematic trends and gaps in academic production in the field, which constitutes a relevant contribution for future research.

During the analysis, conceptual elements necessary to clearly define the object of study were also considered. It was assumed that these are formal research products that, through a rigorous and methodical process, seek to answer a relevant question in an area of knowledge, structuring arguments and evidence that contribute to the academic heritage. For its part, the concept of International Business was understood as the set of commercial activities that cross borders, including transactions of goods, services and foreign direct investment [22,23]. Likewise, it was considered that institutional research policies and the thematic lines defined by each university influence the orientation of academic work, as they establish guidelines, methodological approaches and priority areas for the development of theses [16,24,25].

In short, the combination of official sources, computer tools and systematic screening criteria made it possible to establish a solid and reliable basis for analysing academic output in International Business during the period 2020–2024. The findings derived from this methodology will be presented and discussed in the following section.

3 Results

The RENATI repository, created in 2006 by SUNEDU, centralises the research work submitted for degrees or qualifications at Peruvian universities. Although it was created in 2006, its first thesis publication dates to 2010. This study analyses 2,048 theses in the International Business programme registered in RENATI between 2020 and 2024.

One of the main findings was the high concentration in the assignment of advisors. The 10 most active advisors account for 588 theses, or 28.7% of the total. Each of them supervised between 30 and 87 theses, with an average of 58.8 theses per advisor. Although none exceeds 5% of the total, the most productive advisor supervised 4.25%. This pattern of concentration is consistent with Lotka's Law, where a few authors concentrate most of the production. The

Herfindahl-Hirschman Index (HHI ≈ 1,088) and the Gini coefficient (0.46) confirm an unequal distribution of tutoring. This suggests the need for institutional strategies that redistribute the academic workload, strengthen the participation of advisors with lower demand, and promote more equitable quality supervision (see Fig. 1).

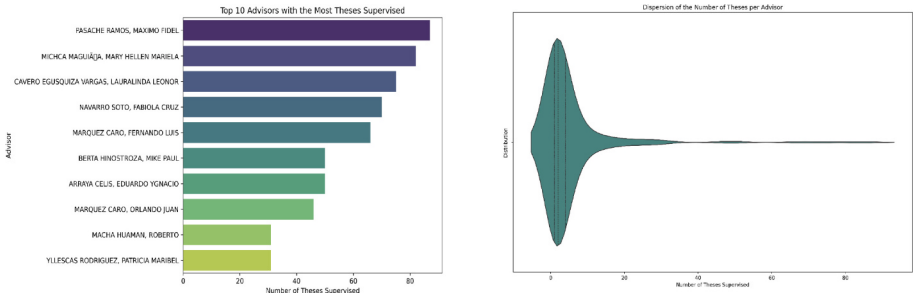


Fig. 1. (A) Top 10 advisors by number of theses. (B) Dispersion of the number of theses per advisor.

The dispersion of the workload among advisors is also shown in Fig. 1(B), which shows that most have between 1 and 8 theses supervised, with a median of around 4–5. However, there are several outliers above 30, 50 and even 80 theses, which generates a pronounced positive asymmetry. This situation combines teacher overload and underutilisation, making it essential to review the tutoring distribution policy, establish reasonable limits and promote the training of new mentors.

Table 1 details the individual workload of the most active advisors, among whom Pasache Ramos (87 theses), Michca Maguiña (82), and Cavero Egusquiza (75) stand out. Together, the three account for almost 43% of the subset analysed. This inequality raises potential bottlenecks that affect academic feedback. International standards recommend not exceeding 10–12 theses per year per advisor, a threshold that is clearly exceeded in the most extreme cases. This could undermine personalised attention, cause delays in completion and affect methodological quality. The institution should monitor these indicators and encourage the participation of new advisors.

Between 2020 and 2024, 2,058 theses were published in RENATI (see Fig. 2), of which 2,048 met the criteria for inclusion in the analysis. The year 2021 stands out as the year with the highest production (588 theses), representing 28.7% of the total. It is followed by 2023 (424), 2020 (407), 2022 (405) and 2024 (224). The annual average is 409.6 theses. The increase in 2020 and 2021 is likely related to the pandemic, which increased the time available to students. The drop in 2024 could be temporary.

Table 1. Top 10 Advisors by Number of Theses (2020-2024)

Advisor Name	Theses
Pasache Ramos, Maximo Fidel	87
Michca Maguia, Mary Hellen Mariela	82
Cavero Egusquiza Vargas, Lauralinda Leonor	75
Navarro Soto, Fabiola Cruz	70
Marquez Caro, Fernando Luis	66
Berta Hinostroza, Mike Paul	50
Arraya Celis, Eduardo Ygnacio	50
Marquez Caro, Orlando Juan	46
Macha Huaman, Roberto	31
Yllescas Rodriguez, Patricia Maribel	31
Average	**58.8**

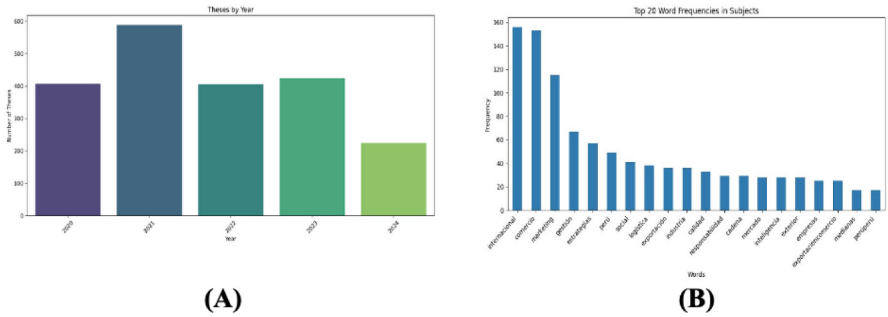


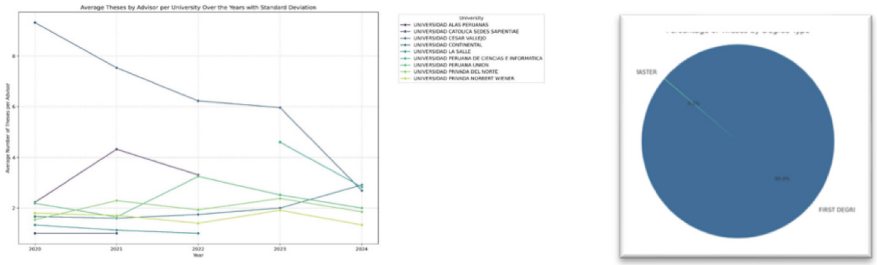
Fig. 2. (A) Number of theses per year. (B) Top 20 most frequently used keywords.

Figures 2(B) and 3 show the analysis of the most frequent keywords. Terms such as "international," "trade," and "marketing" appear more than 150 times. These dominant concepts form a typical pattern of asymmetric distribution, with a "long tail" of less frequent terms. Figure 3, in the form of a word cloud, also reveals the relevance of Peruvian products such as coffee, cocoa, textiles and mangoes, suggesting a strong focus on exports.

The most recurrent terms are "international" (20.87%), "trade" (20.74%) and "marketing" (15.46%). This reflects an academic orientation towards the study of international markets, positioning strategies and commercial management in global contexts. The high frequency of these words indicates the dominant interests among thesis writers and helps to understand the educational trends in the area.

Fig. 3. Word cloud analysis of most frequently repeated keywords in the theses.

In terms of institutional concentration, Fig. 4(A) shows that César Vallejo University leads with 826 theses, followed by the Universidad Privada del Norte (486) and the Universidad Peruana Unión (264). These three universities account for more than 75% of total production, while institutions such as the Universidad Católica Sedes Sapientiae report only two theses. This asymmetry highlights unequal research capacity among universities, which should be considered in institutional improvement and resource distribution policies.



(A) **(B)**

Fig. 4. (A) Number of theses per advisor in the International Business programme from 2020 to 2024. (B) Number of theses by type of degree for the years 2020 and 2024.

In terms of degree type, analysis of the records shows that 99.9% of theses correspond to bachelor's degree programmes, with only three master's theses in the entire period (see Fig. 4(B)). This concentration at the undergraduate level reflects the mandatory nature of the thesis as a requirement for obtaining a professional degree, compared to lower requirements or methodological alternatives at the postgraduate level.

Figure 5 shows the 10 most active jurors in thesis evaluation. Pasache Ramos tops the list with 264 participations, followed by Michca Maguiña (255) and Márquez Caro (227). Together, these three jurors evaluated nearly 55% of the top 10 theses. The group average is 112.6 theses, with a standard deviation of 56.1. The strong concentration of evaluation functions can lead to overload and limit thematic and methodological diversity in research.

Figure 6 illustrates a filtered word co-occurrence network graph, visualizing the top 0.1% of co-occurrence edges among terms extracted from the corpus. In this network, nodes represent individual terms, while edges denote significant co-occurrence relationships between them. Node size is proportional to the degree of connectivity, indicating the relative importance or centrality of each term within the network. Among the most connected nodes are "strategies", "logistics", "company", "market", "product" and "Lima". These links suggest a focus on business planning, foreign trade and regional analysis.

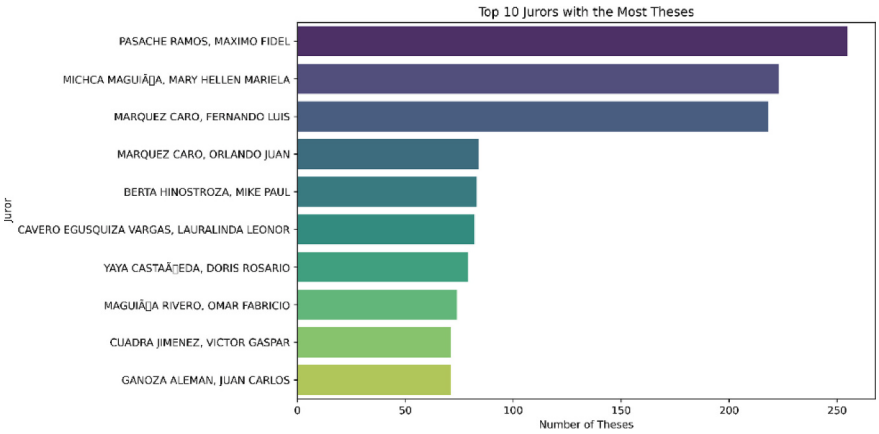


Fig. 5. Top 10 jurors with the most theses evaluated.

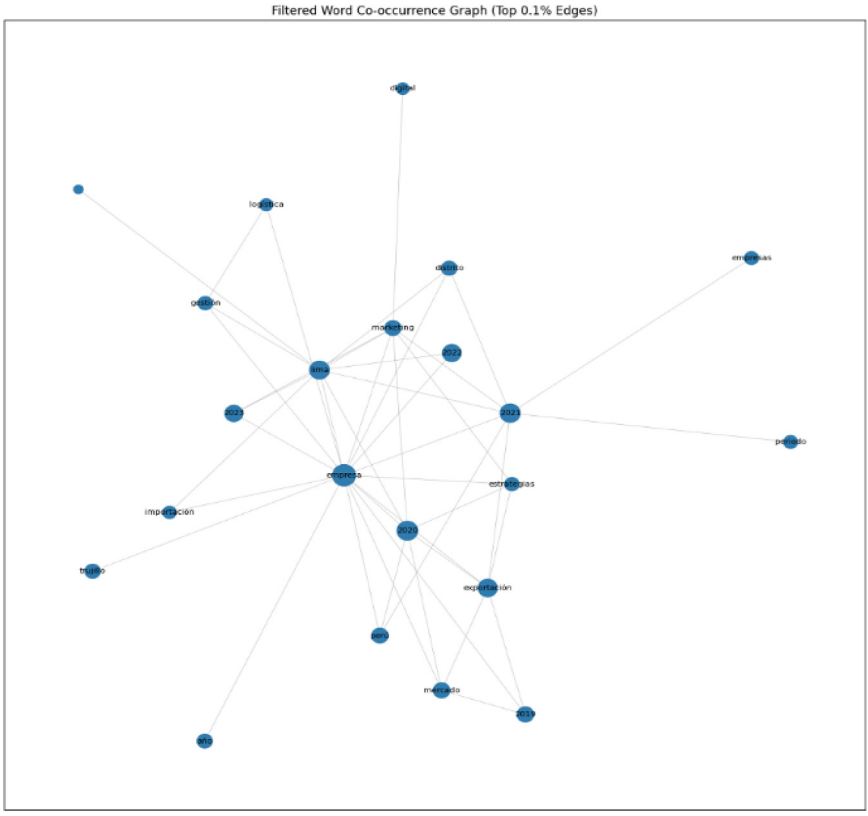


Fig. 6. Co-occurrence network graph highlighting the centrality of topics such as trade, marketing and exports, reflecting the growing relevance of digital transformation and sustainability in international trade during 2020–2024.

The thematic evolution of the last five years is presented in Table 2. Using the LDA (Latent Dirichlet Allocation) model, five major dominant themes were identified: (1) international marketing and global positioning, (2) logistics and supply chain, (3) sustainability and social responsibility, (4) digital transformation and consumer behaviour, and (5) exports and agribusiness.

Table 2. Evolution of the Most Researched Topics (2020-2024)

#	2020	2021	2022	2023	2024
#01	International marketing and quality in Peruvian foreign trade	Import planning and commercial processes in medium-sized companies	Export of agro-industrial products and digital positioning	Sustainable commercial strategies and brand development	International business management and customer experience
#02	Economic growth and export processes in Peru	Digital management and marketing in international trade	Business leadership and free trade agreements	Logistics, costs, and customer satisfaction in foreign markets	Service quality and organisational culture in international sales
#03	Business intelligence strategies and foreign finance	Logistics management and business supply chain	Customer loyalty and digital organisational growth	Consumer behaviour and digital export processes	Strategic marketing and international digital transformation
#04	Consumer analysis and post-COVID logistics	Textile exports and business technology transformation	Export chain strategies and international planning	Artisanal talent and customer behaviour in global markets	Corporate image and social responsibility in exporting
#05	Export planning in medium-sized Peruvian companies	International trade assessment and positioned marketing	Digital marketing and import logistics processes	Strategic management of logistics chains and commercial cargo	International distribution and human change in foreign trade

4 Discussion

The analysis of 2,058 International Business theses registered with RENATI between 2020 and 2024 confirms that Peru is undergoing a process of massification like that of other Latin American countries, where rapid growth in university

enrolment has led to a quantitative expansion in the production of degree theses that has not always been accompanied by quality assurance mechanisms [26,27]. The Peruvian rate is equivalent to an annual average of 412 theses, a figure that places it in the upper range of regional productivity. However, as the literature warns, volume alone does not guarantee improvements in relevance, originality or methodological rigour [28].

Initial evidence of this imbalance can be seen in the distribution by academic level: 99.9% of theses correspond to undergraduate studies, and only three to postgraduate studies. This data suggests that, in the absence of funding and specialised tutoring, research is relegated to the initial cycles, truncating the training chain and weakening the generation of advanced knowledge [29–31]. As a result, the local academic contribution to the international debate on international business is mainly based on studies of limited scope, with little theoretical or methodological projection.

Another critical feature is the concentration of mentoring. A small core of advisors accounts for more than 40% of thesis supervision, with some advisors supervising more than 80 theses in five years. Comparative studies have shown that overloading thesis supervisors reduces the quality of supervision, increases completion times and raises the rate of substantial revisions [32–35]. In the Peruvian case, it has been observed that the most sought-after advisors tend to use standardised research templates and prefer descriptive designs that can be implemented quickly.

In terms of content, the most frequent keywords (commerce, marketing, logistics) reveal a preference for traditional topics. In contrast, terms such as e-commerce, sustainability, and data analytics barely exceed 5% frequency. This gap coincides with findings in other regions, where a disconnect has been documented between the topics chosen by students and emerging priorities linked to digitalisation and Industry 4.0.

A complementary tool used in this study was keyword co-occurrence analysis. Unlike simple frequency counting, this technique allows for the visualisation of thematic relationships between concepts that appear together in theses. Dense conceptual clusters were identified—such as the association between "logistics" and "customer," or between "trade" and "digital"—which reveal already established lines of research.

The thematic evolution between 2020 and 2024 shows a gradual transition from functional approaches, focused on marketing and foreign trade, to more integrative themes such as customer experience, digital transformation and social responsibility. This shift reflects greater alignment with international trends and suggests a slow but positive move towards diversification in the field.

5 Conclusions

The analysis of 2,058 International Business theses registered in RENATI between 2020 and 2024 reveals an unprecedented expansion in undergraduate research production following the implementation of Law 30220. This growth

confirms the effectiveness of mandatory and open access mechanisms, but also exposes a model based on the mass reproduction of research that does not always guarantee methodological soundness or academic impact.

The virtual absence of postgraduate theses—only three in five years—reflects a structural gap that limits advanced research. The low participation of master's and doctoral programmes suggests that theses continue to be treated as an initial qualification requirement, disconnected from a sustained research career.

Another critical problem is the concentration of tutoring: less than ten advisors account for more than 40% of all supervision. This overload leads to the use of homogeneous templates, limited personalised feedback and less exploration of sophisticated theoretical or methodological frameworks.

In terms of content, lexical analysis of titles and keywords reveals a persistent thematic lag. Trade, marketing and logistics continue to be the dominant themes, while key concepts in the contemporary economy—such as sustainability, digitalisation or data analytics—barely appear.

Given this situation, it is recommended that institutions:

- Limit the number of theses per supervisor and encourage specialised co-supervision
- Modernise the International Business curriculum, incorporating compulsory seminars on digital commerce, sustainability, data analytics and applied artificial intelligence
- Strengthen postgraduate programmes with competitive scholarships and access to specialised databases
- Optimize RENATI governance through standardisation of metadata and anti-plagiarism controls

Finally, future research should evaluate the content of theses and their subsequent impact, identifying how many are translated into publications, patents or public policies.

Acknowledgments. The authors would like to thank the Research Directorate of the Universidad Peruana de Ciencias Aplicadas, which provided financial support through UPC-EXPOST2026.

Disclosure of Interests. No potential conflict of interest was reported by the author(s).

References

1. Hofmann, P.: The impact of international trade and FDI on economic growth and technological change (2013)
2. Moon, B.E.: Dilemmas of international trade, 2nd edn (2018)
3. Angeles-Castro, G., Perrotini-Hernández, I., Ríos-Bolívar, H.: Market liberalism, growth, and economic development in Latin America (2011)
4. U.S. Department of State: Economic Policy and Trade Practices Report: Peru. Washington, DC (1996)

5. World Bank: Trade (% of GDP) – Peru, 1990–2024 (2025)
6. APEC Secretariat: Peru: Member Economy Profile (2023)
7. World Trade Organization: Peru and the WTO: Trade Policy Review (2023)
8. World Bank: Peru – Trade and tariff indicators (1990–1999) (1999)
9. United Nations Conference on Trade and Development: World Investment Report 2025: Annex Table 01 – FDI inflows by economy, 1990–2024 (2025)
10. Caraza, M.G.: Peru and international trade: a global view. Manag. Third Millennium **2**, 81–90 (1999). https://doi.org/10.15381/gtm.v2i3.10050
11. Varona-Castillo, L., Gonzales-Castillo, J.R.: Economic growth and income distribution in Peru | Crecimiento económico y distribución del ingreso en Perú. Probl Desarro **52**, 79–107 (2021). https://doi.org/10.22201/iiec.20078951e.2021.205.69636
12. Kaynak, E., Schermerhorn, J.R.: Teaching and programme variations in international business (2021)
13. RENATI: Statistics on registered theses (2024)
14. Congress of the Republic of Peru: University Law (2014). https://www.sunedu.gob.pe/wp-content/uploads/2017/04/Ley-universitaria-30220.pdf, Peru
15. Masías, L.: SUNEDU presents digital repository for research papers (2025)
16. Arcila-Diaz, J., Delgado-Caramutti, J., Millones-Gómez, P.A., et al.: Research trends in Peruvian universities: proposal for a research agenda with a bibliometric approach. Scientometrics **130**, 489–514 (2025). https://doi.org/10.1007/s11192-024-05211-z
17. Alvan Romero, R.M.A., Castro Romero, J.: Evaluation of the main internal barriers limiting cocoa exports, with a view to improving the social conditions of farmers in the community of Constitución in the department of Pasco between 2017 and 2021 (2023)
18. Bautista Urbano, N., Cárdenas Paz, R.G.N.: Analysis of the main constraints faced by Peruvian companies in exporting software in the period 2015-2019
19. Gambini Aquino, C.A.: International trade barriers and business management among workers in a micro-enterprise in northern Lima (2024)
20. Cantillo Velásquez, I.M., León Vargas, Y.: Research trends in international business. Trends **23**, 273–305 (2022). https://doi.org/10.22267/rtend.222302.209
21. Georgia Southern University: Dissertations & Theses: Understanding Embargoes (2020). https://georgiasouthern.libguides.com/c.php?g=834055&p=6939131. Accessed 22 June 2025
22. Fernández, J.: International business: what is it and why is it important? (2024)
23. Guillén, M.F.: International business. In: International Encyclopedia of the Social & Behavioral Sciences, pp. 480–483. Elsevier (2015)
24. Campo, H.: What is a research policy and why is it important for a country's development? (2019)
25. Andrés Bello University: Research policies (2023)
26. Mendoza, P., Dorner, L.: The neoliberal discourse in Latin American higher education: a call for national development and tighter government control | O discurso neoliberal na educação superior latino-americana: Uma chamada para o desenvolvimento nacional e um controle governamen. Educ Policy Anal. Arch. **28**, 1–28 (2020). https://doi.org/10.14507/epaa.28.5610
27. Sumar Albujar, O.A., Orellana Presentación, J.C., Estrada, R.H.: Escaping the "Regulatory State": the case of the peruvian tertiary education regulations | Escapando del "Estado Regulatorio": el caso de la regulación de la educación superior en Perú. Euro-Latin Am. J. Adm. Law **10** (2023). https://doi.org/10.14409/redoeda.v10i1.12759

28. Forthmann, B., Leveling, M., Dong, Y., Dumas, D.: Investigating the quantity-quality relationship in scientific creativity: an empirical examination of expected residual variance and the tilted funnel hypothesis. Scientometrics **124**, 2497–2518 (2020). https://doi.org/10.1007/s11192-020-03571-w
29. Beech, J., López, C.A.: The development of comparative education in Latin America: biographies, histories and transnational forces (2025)
30. Palamidessi, M.I., Gorostiaga, J.M., Suasnábar, C.: The development of educational research and its links with the government of education in Latin America. Educ. Profiles **36**, 49–66 (2014). https://doi.org/10.1016/S0185-2698(14)70609-9
31. Guzman-Valenzuela, C.: Global trends in higher education and their impact on Latin America: Pending challenges | Global trends in higher education and their impact on Latin America: Pending challenges. Lenguas Modernas **2017**, 15–32 (2017)
32. Abdullah, N., Sherrif, N.M.: Effect of supervisor quality and teaching quality on students' satisfaction and loyalty in Malaysian Research University. Adv. Nat. Appl. Sci. **6**, 1428–1437 (2012)
33. Bayona-Oré, S.: Supervisor performance and thesis development in postgraduate students. In: ACM International Conference Proceeding Series, pp. 109–113 (2018)
34. Gupta, S., Sharma, J.: Evaluating research students' satisfaction based on supervisors' characteristics: a fuzzy approach. In: Proceedings of the 2013 IEEE International Conference in MOOC Innovation and Technology in Education Mite 2013, pp. 190–195 (2013)
35. Ringo, D.S.: When relationships make a difference! Supervisor support and PhD students' academic performance in the journey: the moderating effect of supervisor-student relationship quality. Cogent Educ. **12** (2025). https://doi.org/10.1080/2331186X.2025.2479403

Clustering Analysis of Student Satisfaction–Performance Relationships in Higher Education Using Large-Scale Educational Data

Michael Newton Cutipa-Santi(✉), Daisy Crishyely Quelca-Velasquez, Yeni Cutipa-Santi, Renzo Apaza-Cutipa, and Vladimiro Ibañez-Quispe

Universidad Nacional del Altiplano, Puno, Peru
{michael.cutipa,renzo,vibanez}@unap.edu.pe

Abstract. This study explores the relationship between student satisfaction and academic performance using advanced clustering techniques on a large dataset from the National University of Altiplano in Puno (UNAP), which includes over 18,000 students. A representative sample of 12,156 students was selected. The analysis focused on 32 items related to student satisfaction, covering teaching quality and university services, which were combined with students' semester averages and admission scores. The data were managed using a preprocessing approach that included normalization and data cleaning, followed by clustering techniques based on the K-Means algorithm. Dimensionality reduction was performed through Principal Component Analysis (PCA) to enhance data interpretation and facilitate the identification of relevant patterns. The elbow method and the silhouette coefficient indicated that the optimal number of clusters was three (K=3). The results revealed three student profiles: Cluster 0 (C0), which includes students with low performance (12.96) and high satisfaction (teaching: 1.87; services: 1.69); Cluster 1 (C1), which groups students with high performance (17.05) and low satisfaction (teaching: 0.94; services: 0.78); and Cluster 2 (C2), composed of students with average performance (15.09) and high satisfaction (teaching: 1.88; services: 1.68). Additionally, a small group of 146 students with low performance and low satisfaction was found within Cluster 1. The validation of the findings through Analysis of Variance (ANOVA) revealed statistically significant differences between the groups ($p<0.001$), reinforcing the validity of the identified profiles. This work provides valuable tools for educational segmentation that can be used to design differentiated and personalized academic policies.

Keywords: Academic Performance · Clustering · Educational Big Data · K-Means · Student Satisfaction

1 Introduction

The connection between student satisfaction and academic performance has attracted increasing attention, as it directly influences retention, achievement,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 433–449, 2026.
https://doi.org/10.1007/978-3-032-20322-9_30

and success in universities. According to UNESCO (2021), the dropout rate in Latin American universities reaches up to 52%. In the current context of higher education, assessing student satisfaction has become a key tool for measuring educational quality and the effectiveness of institutional processes. This satisfaction reflects how students perceive various aspects of their academic environment, such as the quality of teaching, the availability of services, and institutional infrastructure [6,24,29].

The study of the relationship between student satisfaction and academic performance has been addressed in several previous studies. However, many of these studies have focused on examining one of these variables in isolation. For example, there are studies that have focused solely on satisfaction without considering how it relates to academic performance [7,8,25]. Other works have explored academic performance without considering the effect that students' perceptions of their educational environment can have, which has complicated a more comprehensive understanding of student success.

However, to obtain a more comprehensive view of educational quality, it is essential to link subjective perceptions with objective indicators such as academic performance [14]. The relationship between student satisfaction and academic performance offers a deeper understanding of the factors that influence student success, allowing universities to implement data-driven improvements [25].

Although universities collect large volumes of data from various sources such as satisfaction surveys, academic records, and socio-economic information, they often fail to leverage it effectively. Satisfaction and academic performance measurements are often conducted in isolation, without exploring the connections between them, which limits the development of effective strategies for continuous improvement [5,8]. The lack of data integration hinders a deeper understanding of the factors affecting student performance and complicates informed decision-making within institutions [7,13].

Despite the abundance of collected data, universities still face challenges in integrating and effectively analyzing student satisfaction metrics and academic performance [28,31]. The absence of correlation between these measurements prevents a complete understanding of the factors influencing student outcomes. This study addresses this gap by combining subjective and objective variables using advanced segmentation techniques such as K-Means and PCA. By identifying three distinct student profiles, this work offers a deeper understanding of variations in student behavior and provides a framework for creating more effective educational policies. Segmenting students based on their satisfaction and academic performance represents an innovative approach that not only improves data interpretation but also enables more personalized interventions, focusing on the groups that need it the most.

The proposed work is based on a hybrid K-Means, PCA model that enhances robustness and interpretability. Evidence from large-scale data for a Latin American public university is presented, translating the profiles into concrete institutional actions. It is framed within a hypothetical integrative binomial model, using unsupervised clustering techniques to identify hidden student profiles based

on nested satisfaction survey responses and objective academic information resulting from students' decisions: grades, admission scores, etc. The National University of Altiplano in Puno (UNAP) can be considered a representative case study of public institutions with significant social diversity. By combining subjective and objective variables, the goal of this approach is to access more information on the criteria that determine academic success and pave the way for the continuous improvement of public universities.

2 Methodology

2.1 Clustering Analysis in Educational Data

Educational data analysis has greatly benefited from the application of clustering algorithms that adapt and function as tools to identify patterns, segment large volumes of information, and gain a deeper understanding of student behavior [34]. The analysis examines various clustering techniques, their specific applications in the educational context, the relationship between satisfaction and academic performance, and the inherent challenges of working with heterogeneous and large-scale data [27,31].

2.2 Fundamentals and Clustering Analysis Techniques

Classical Algorithms: K-Means, Hierarchical, and DBSCAN. The K-Means algorithm has become a fundamental tool in educational data mining. Its simplicity and efficiency are useful for categorizing students into homogeneous groups with similar characteristics, which facilitates the identification of meaningful patterns in academic performance and satisfaction [2,32]. Hierarchical clustering, on the other hand, opens a range of possibilities for group analysis. A notable model is HKMC, which combines K-Means with hierarchical dendrograms to identify patterns among students.

In contrast, the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) has been applied in various studies. Its ability to identify clusters of any shape, even non-spherical, is advantageous for detecting complex patterns [35]. However, DBSCAN has limitations when working with homogeneous data, such as admission scores, due to the difficulty of fine-tuning distance and density parameters. Despite this, it remains a valuable tool for educational data analysis, where the shape and density of clusters can vary significantly.

Advanced Approaches: Fuzzy and Hybrid Clustering. Fuzzy clustering techniques have gained relevance in educational data analysis. These allow students to belong to multiple groups with different degrees of membership, which is useful when dealing with complex and multidimensional characteristics [16,17]. Hybrid approaches, which combine clustering techniques with neural networks or deep learning models, allow for predicting future student performance and generating personalized educational interventions [17,30].

The combination of K-Means, hierarchical clustering, and PCA was a strategy to leverage the best of each method in clustering analysis. K-Means was chosen for its ability to segment large volumes of data into well-defined clusters, which is essential for identifying student profiles. On the other hand, hierarchical clustering, using the Ward method, was key to estimating the optimal number of clusters, which led to deciding on K=3 for K-Means. Finally, PCA helped reduce the dimensionality of correlated variables, improving cluster visualization and confirming that the identified groups were consistent with the underlying structure of the data.

2.3 Dimensionality Reduction: PCA (Principal Component Analysis)

Principal Component Analysis (PCA) is an unsupervised technique that works with large datasets. PCA reduces the dimensionality of the data without losing the essential variability needed for segmentation, which makes algorithms like K-Means work more effectively [24].

2.4 Applications of Clustering in Education

Clustering has been widely used in the educational field to segment students based on various characteristics, such as their grades, satisfaction with institutional services, or behavior in the classroom [21]. Segmentation allow universities to identify groups with similar academic needs and create personalized support programs [10].

Educational segmentation through clustering is crucial for improving the understanding of student profiles, enabling the design of more effective educational policies. Research has shown that proper segmentation can lead to significant improvements in student retention and academic performance, especially when subjective factors (such as satisfaction with teachers and university services) are combined with objective metrics [15]. This personalization is key to improving performance and optimizing resources by targeting efforts where they are most needed [29].

2.5 Relationship Between Student Satisfaction and Academic Performance

The relationship between student satisfaction and academic performance has been the subject of various studies with mixed results. Some suggest that higher satisfaction leads to better performance due to increased motivation and engagement [23]. However, more recent studies indicate that satisfaction does not directly translate to better performance, as factors like stress, personal difficulties, or emotional issues can affect the performance of satisfied students [36].

Nachouki and Naaj [20] propose that satisfaction and performance should be considered separately and combined with other metrics, such as admission

scores, to gain a more comprehensive view of academic success. This approach helps address the paradox of high satisfaction but low performance and identify strategies to improve both aspects in specific educational contexts.

2.6 Processing Large Volumes of Data

The analysis of large volumes of educational data, including academic histories, assessments, interactions on virtual platforms, attendance data, or performance metrics, requires processing capabilities that exceed a single computer or server [33]. Distributed computing emerges as a solution, as it divides tasks across multiple nodes or machines working in parallel, significantly reducing execution time and enabling the handling of datasets that exceed the limits of local memory [4].

Platforms like Apache Spark have revolutionized the field by providing an optimized framework for massive in-memory data processing, supporting more complex tasks like machine learning and advanced statistical analysis [1]. Thanks to technological advances, researchers scale clustering algorithms like K-Means or DBSCAN to millions of records, facilitating the segmentation of student populations and the discovery of learning patterns, while maintaining the necessary efficiency and responsiveness for data-driven decision-making in large-scale educational environments [18].

2.7 Statistical Validation and Future Challenges

Metrics such as Analysis of Variance (ANOVA), the silhouette coefficient, and the elbow method are used to validate clustering results and determine the optimal number of clusters. Additionally, performance metrics like accuracy, recall, and F1-score are essential for measuring the quality of the generated clusters [10].

Challenges of Data Heterogeneity. Despite advances, the heterogeneity of student data (behavior, performance, socio-economic context, mental health) represents a significant barrier. Models must be capable of handling massive information for processing [9].

Impact on the Design of Academic Policies. Cluster analysis provides a better understanding of the factors influencing performance and satisfaction, facilitating the creation of personalized academic interventions [11].

2.8 Clustering and PCA

To ensure the robustness of the results, a hybrid approach was adopted, combining the robustness of multiple techniques in a logical sequence. First, the variables admission score, teaching index, services index, and semester averages were standardized, ensuring that all variables had equal influence on the analysis. Dimensionality reduction was then carried out using Principal Component

Analysis (PCA) to facilitate exploratory visualization. Two principal components were retained, explaining 58.4% of the total variance, simplifying the dataset, eliminating multicollinearity and noise, and improving the quality of the clusters.

Hierarchical clustering was then used as an exploratory technique to determine the optimal number of clusters. This method mitigated one of the main limitations of the K-Means algorithm, which is the difficulty in selecting the appropriate value of K without objective guidance. Finally, with the validated optimal number of clusters (K=3), the K-Means algorithm was applied for final segmentation. The choice of K-Means was based on its computational efficiency and ability to handle large volumes of data quickly and accurately, resulting in well-defined and easily interpretable segments.

The differences between the identified clusters were validated through one-way Analysis of Variance (ANOVA) ($p<0.001$), and post-hoc tests (Tukey HSD) were performed to confirm the significance of the differences.

3 Results

3.1 Number of Clusters

The analysis revealed that the model's inertia decreased significantly to $K = 3$, dropping from 85,092 to 66,386, indicating a considerable improvement in the cohesion of the groups. Additionally, the silhouette coefficient reached an average of 0.12 for $K = 3$, which is acceptable and indicates moderate separation between the clusters, reflecting good internal cohesion along with adequate external differentiation. This result validates the selection of $K = 3$ as the optimal number of clusters.

3.2 Projection and Visualization with PCA

Based on the analysis of the means of the key variables (see Table 1), the following cluster profiles were identified:

Cluster 0 (C0): Students with low academic performance (12.96) and high satisfaction. This group is characterized by having the lowest academic average, despite its highest satisfaction indices (teaching: 1.87; services: 1.69).

Cluster 1 (C1): Students with high academic performance (17.05) and low satisfaction. This group shows the highest academic average, but the lowest satisfaction indices (teaching: 0.94; services: 0.78).

Cluster 2 (C2): Students with average performance (15.09) and high satisfaction. This cluster shows average academic performance with high satisfaction indices (teaching: 1.88; services: 1.68).

Using PCA to project the data, a clear separation between the three main groups was observed. The distribution was reflected in a color-coded scatter plot, which helped provide a better visualization of the cluster structure and confirmed that the model matched what was observed in the reduced space.

Figure 1 presents a two-dimensional projection of the data obtained through Principal Component Analysis (PCA). Each data point represents a student and

Table 1. Summary of Key Variables by Cluster (Mean and Standard Deviation)

Cluster	Admission	Teaching Index	Services Index	Semester Avg. 2024-II
C0	2417.98	1.87	1.69	12.98
C1	2437.60	0.94	0.78	17.02
C2	2461.24	1.88	1.68	15.10

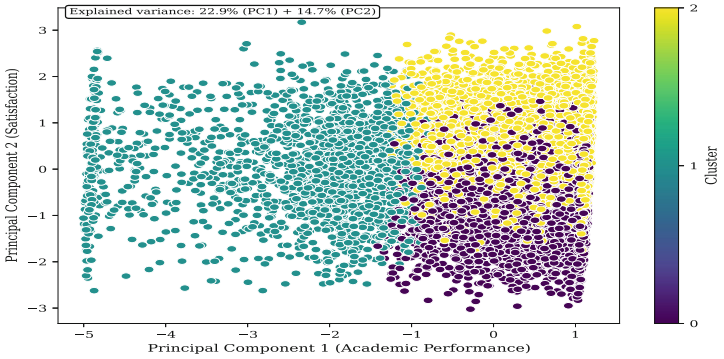


Fig. 1. Student Segmentation Using K-Means Clustering ($K = 3$) Projected in PCA Space. Each point represents a student; the colors indicate cluster membership.

has been colored according to the K-Means clustering solution ($K = 3$). The visualization clearly shows a separation between the three groups, suggesting that the clustering model has identified distinct student profiles.

The separation of the groups occurs primarily along the First Principal Component (PC1), which captures most of the variance in the dataset. In this context, PC1 can be interpreted as an academic performance axis, where lower values correspond to lower performance and higher values to better performance. The Second Principal Component (PC2), with less explained variance, is related to student satisfaction, serving as a secondary differentiator.

Qualitatively, the profiles of the clusters are distributed as follows in this space:

- **Cluster C0** is concentrated at lower values of PC1, indicating lower academic performance. This group is particularly notable because, according to subsequent analyses, it shows high satisfaction despite its low performance.
- **Cluster C1** occupies a range of intermediate to high PC1 values, with comparatively lower values in PC2, suggesting a profile of medium to high academic performance with low satisfaction.
- **Cluster C2** is positioned in an intermediate to high range of PC1, with higher values in PC2, associated with medium to high academic performance and high satisfaction.

This pattern of segmentation visually confirms that the clusters are not random but instead reflect well-defined underlying profiles in the student popula-

tion, which align with the performance and satisfaction metrics presented in the tables.

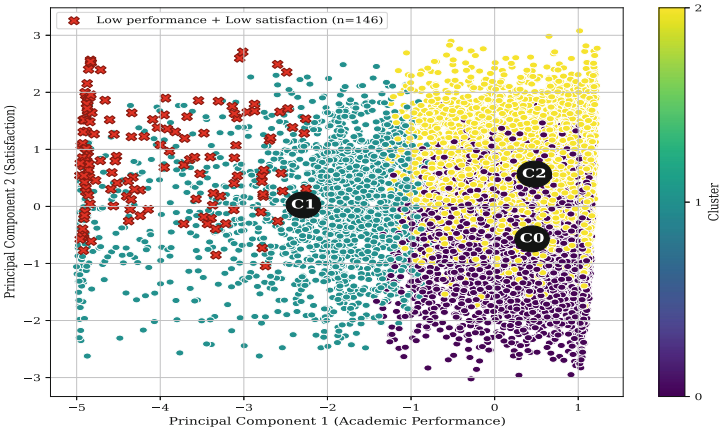


Fig. 2. Student Segmentation Using Hierarchical Clustering (Ward's Method, $K = 3$) Projected in PCA Space.

Figure 2 presents a two-dimensional projection of the students obtained through Principal Component Analysis (PCA) of the standardized variables. The points are colored according to the hierarchical clustering solution using Ward's method with $K = 3$. The distribution of clusters along the first principal component (PC1) and overlap in the second principal component (PC2) is similar to that obtained with K-Means (Fig. 1), suggesting that the clustering is primarily driven by academic performance (PC1) and, to a lesser extent, by student satisfaction (PC2).

From the visualization, the profiles of the resulting clusters are as follows:

- **Cluster C0** is located at intermediate-high values of PC1, with comparatively lower values in PC2, suggesting medium to high academic performance with relatively low satisfaction.
- **Cluster C1** is situated at lower values of PC1 and high values in PC2, indicating low academic performance with high satisfaction.
- **Cluster C2** is also located at intermediate-high values of PC1, but with higher values in PC2, suggesting medium to high academic performance with relatively high satisfaction.

Additionally, the figure highlights the 146 students with low performance and low satisfaction, marked with a red icon. Their position at the lower-left corner of the plot, far from the cluster centers, shows that they do not form a cohesive group on their own. Instead, they are primarily dispersed within Cluster C1, demonstrating the robustness of the three-cluster model by not creating an additional group from these outliers.

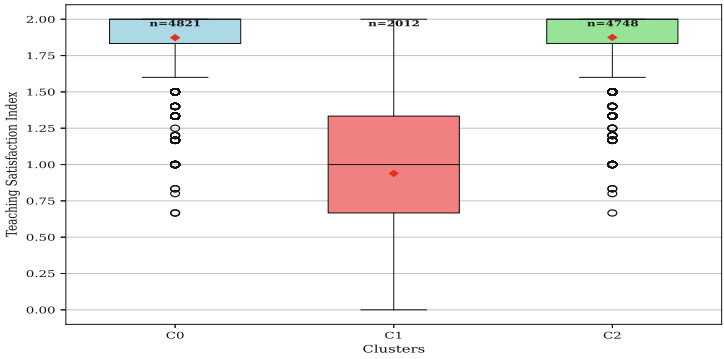


Fig. 3. Distribution of Teaching Satisfaction Index by Cluster ($K = 3$).

Figure 3 shows the distribution of the teaching satisfaction index for the three identified clusters, using a box plot. Clusters C1 and C2 show consistently high satisfaction, with medians close to 2.00 and interquartile ranges between 1.75 and 2.00, reflecting generally favorable perceptions of teaching quality. Both clusters have few outliers below 1.5, which may reflect isolated cases of dissatisfaction within an otherwise favorable group.

In contrast, cluster C0 shows a considerably lower median (1.00) and greater dispersion (0.00–2.00), suggesting more heterogeneity in the evaluations. This cluster has a higher proportion of negative responses, indicating a less favorable perception of teaching quality compared to the other clusters.

The observed differences between the clusters highlight that satisfaction with teaching acts as a strong differentiator between the groups. Additionally, the low satisfaction outliers in C1 and C2 reflect individual cases that warrant qualitative follow-up to identify potential factors that could lead to student disengagement.

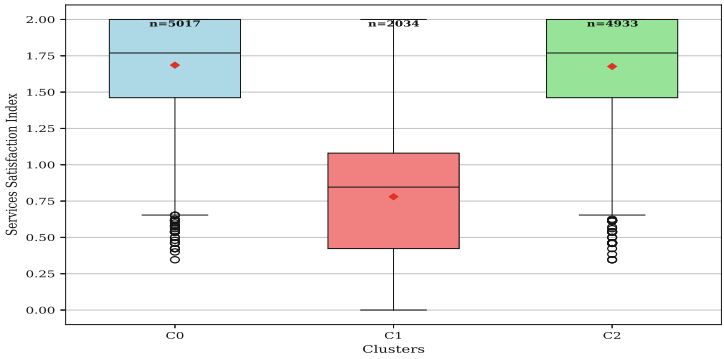


Fig. 4. Distribution of Service Satisfaction Index by Cluster ($K = 3$).

Figure 4 presents the distribution of the University Services Satisfaction Index for the three clusters identified through K-Means ($K = 3$).

Cluster C2 shows consistently high satisfaction, with a median close to 1.75 and a narrow interquartile range (1.50–2.00), reflecting consistent and favorable perceptions of university services. In contrast, clusters C0 and C1 show markedly lower medians, around 0.83, and a wider dispersion (0.00–2.00), indicating greater heterogeneity in perceptions and a higher proportion of negative evaluations.

The observed pattern reinforces the role of perceived service quality as a key differentiator in segmentation. The low satisfaction outliers in cluster C2 reflect individual cases that deserve qualitative follow-up to identify potential factors that could lead to student disengagement.

Interpretation of Boxplot Elements:

- **Median Line:** Central tendency of satisfaction scores for each cluster.
- **Box Limits:** Interquartile range (IQR), which captures the central 50% of responses.
- **Whiskers:** Range of most values, excluding outliers.
- **Outliers:** Individual cases with extreme deviations from the central tendency of the cluster.

Overall, Fig. 4 indicates that, while C1 and C2 maintain uniformly positive evaluations of the services, the small proportion of dissatisfied students within them deserves follow-up. In contrast, the lower median and greater variability of C0 highlight the need for specific improvements in service delivery processes.

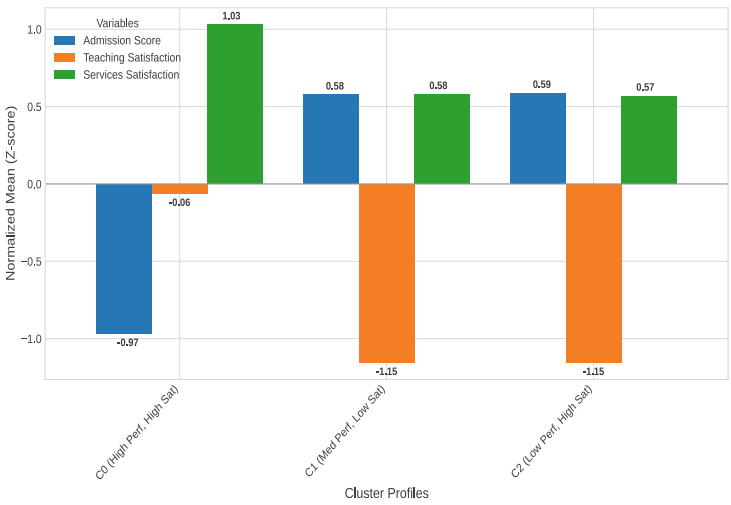


Fig. 5. Comparison of Mean Values and Dispersion (Standard Deviation) of Key Variables.

Figure 5 compares the mean values and dispersion (standard deviation) of key variables: admission score, teaching satisfaction index, and service satisfaction index across the three identified clusters ($K = 3$). Cluster C0 (High Performance, High Satisfaction) stands out for having consistently high normalized values for both teaching satisfaction and service satisfaction. In contrast, cluster C1 (Medium Performance, Low Satisfaction) shows much lower levels in both areas. Cluster C2 (Low Performance, High Satisfaction) also maintains high satisfaction levels.

Regarding admission scores, the profiles of clusters C1 and C2 are quite similar, while cluster C0 has a slightly lower score. This pattern suggests that, although there are no significant differences in admission performance, perceptions of teaching quality and services are key factors that help differentiate the groups.

Although the results revealed three main clusters, a group of 146 students with low performance and low satisfaction was also identified. These students were integrated into the existing clusters. This finding represents a profile with very specific needs. For future research, it would be advisable to conduct a more detailed analysis of the students in this category to better understand their characteristics and design tailored interventions for this group.

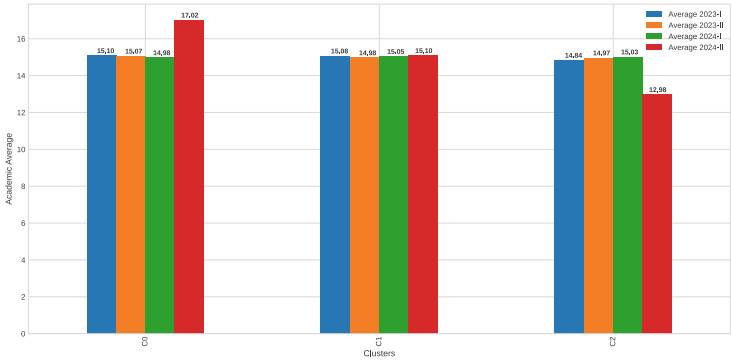


Fig. 6. Average Academic Performance over Four Consecutive Semesters.

Figure 6 shows the average academic performance over four consecutive semesters (2023-I, 2023-II, 2024-I, and 2024-II) for the three clusters ($K = 3$).

Cluster 0 (C0): This cluster maintains stable performance around 15 points in all periods.

Cluster 1 (C1): This group consistently shows the highest averages, with a marked increase in the semester 2024-II, surpassing 16 points.

Cluster 2 (C2): This cluster maintains a level similar to C0 during the first three semesters but stands out in 2024-II with an average that does not exceed 14 points.

The observed trend is clear: Cluster 1 shows a notable improvement in performance, while Cluster 2 is experiencing a slight decline. This difference in academic performance is a key factor in understanding how these groups are segmented.

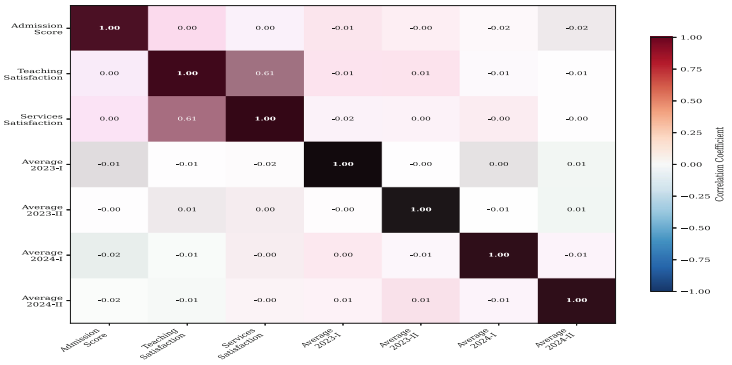


Fig. 7. Correlation Matrix Between Admission Score, Satisfaction Indices, and Academic Performance.

Figure 7 illustrates the correlation matrix between key variables: admission score, teaching satisfaction index, service satisfaction index, and academic averages for the semesters 2023-I, 2023-II, 2024-I, and 2024-II. The results show a moderate positive correlation, around 0.6, between teaching satisfaction and service satisfaction. This suggests that students who value teaching quality also tend to appreciate the services offered by the institution.

On the other hand, a strong positive correlation, approximately 0.7, is observed between admission score and academic performance across the semesters. In contrast, the correlations between the satisfaction indices (teaching and services) and academic performance are nearly zero, indicating no direct linear relationship between these variables. This finding implies that performance differences are not explained by satisfaction, justifying the need for a clustering analysis to capture more complex interactions.

4 Discussion

The segmentation results obtained through K-Means and hierarchical clustering (Ward's method) consistently identified three groups of students, differentiated primarily by academic performance variables and secondarily by satisfaction indices. This finding is consistent with [26], who reports that clustering methods can effectively characterize distinct student profiles based on academic and perceptual data.

The PCA projections confirmed that the first principal component (PC1) is dominated by admission scores and semester averages (GPA), while the second component (PC2) captures variability in satisfaction with teaching and

university services. This pattern supports the observations of [3], who reports that performance-related variables explained most of the variance in educational datasets, while affective or perceptual variables act as complementary differentiators.

Clusters with higher satisfaction (C0 and C2) tend to maintain stable or improving performance, which is consistent with evidence that data-driven segmentation can reveal profiles linked to differential outcomes and inform targeted interventions [2]. In contrast, the cluster with lower satisfaction (C1) shows greater variability and consistently lower averages, suggesting the presence of contextual or systemic barriers, such as limited access to academic resources or weaker student-teacher interactions, as discussed in [12].

The marked improvement in performance in C1 during the 2024-II semester (Fig. 6) may be associated with specific academic interventions or favorable institutional changes. Previous research, such as [12], has shown that improving service quality and adopting digital support strategies can foster greater engagement and academic achievement—an interpretation that aligns with the observed trajectory of C1. However, the correlation matrix reflected that admission scores and satisfaction indices did not exhibit a strong linear relationship with academic performance, reinforcing the idea that differences were driven by more complex, non-linear interactions, better captured through cluster analysis.

Together, the findings highlight the importance of combining quantitative segmentation with qualitative follow-up to achieve a richer understanding of the nuances of the student experience and to design evidence-based interventions [6]. For group C0 (low performance and high satisfaction), it is recommended to implement intensive academic tutoring along with emotional support and stress management programs. For group C1 (high performance and low satisfaction), efforts should focus on improving service quality, feedback mechanisms, and student-teacher interactions. Finally, for group C2 (average performance and high satisfaction), it is suggested to promote leadership roles and enrichment activities to help sustain and enhance their positive trajectory [35], [?, ?] [19].

The main threats to the validity of this study include response bias in the student satisfaction surveys, which could affect the accuracy of the subjective data. Additionally, the potential non-representativeness of the student sample relative to the general university population limits the generalizability of the results. The selection of the clustering method also presents an inherent limitation, as the algorithm may not fully capture the variability in the data, especially in such a diverse educational context. For future studies, it is recommended to include additional variables and consider more flexible methodological approaches to address these threats to validity.

5 Conclusion

The study showed that combining K-Means and hierarchical clustering with PCA was an effective strategy to identify different student profiles, based on

academic performance and satisfaction indicators. The cluster analysis consistently revealed three groups: Cluster C0: Students with low performance but high satisfaction. Cluster C1: Students with high performance and low satisfaction. Cluster C2: Students with average performance and high satisfaction.

The findings confirmed that academic performance is the main factor driving segmentation, while perceptions of teaching and services are secondary differentiators, though important. These results have direct implications for institutional policy such as: providing academic reinforcement and resource support for group C0, maintaining favorable conditions for group C2, and replicating the successful interventions observed in C1, which shows improvements in performance despite low satisfaction.

Additionally, the analysis highlighted that academic performance is the key factor in segmentation, with satisfaction indices acting as a secondary but significant differentiator. Clusters with high satisfaction (C0 and C2) show stable or improving performance, while the cluster with low satisfaction (C1) presents greater variability and consistently lower averages. The PCA analysis confirmed that admission and performance metrics dominate the first principal component, while satisfaction indicators explain the second. The correlation results revealed weak linear associations between satisfaction/admission and performance, highlighting the importance of non-linear methods such as clustering.

The methodological approach, which combines quantitative segmentation with satisfaction metrics, was solid and can be adapted to other educational contexts.

Acknowledgments. The authors express their gratitude to the National University of Altiplano for granting access to the educational data that made this research possible.

Disclosure of Interests. The authors declare no relevant conflicts of interest with the content of this article.

References

1. Abdrakhmanov, R., Zhaxanova, A., Karatayeva, M., Niyazova, G.Z., Berkimbayev, K., Tuimebayev, A.: Development of a framework for predicting students' academic performance in stem education using machine learning methods. Int. J. Adv. Comput. Sci. Appl. **15**(1) (2024)
2. Acosta, L.D.I., Polo, J.E.R.: Identification of factors that affect the academic performance of high school students in peru through a machine learning algorithm. In: Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology (2021). https://doi.org/10.18687/LACCEI2021.1.1.68
3. Aguilar, S.J.: Examining the relationship between comparative and self-focused academic data visualizations in at-risk college students' academic motivation. J. Res. Technol. Educ. **50**(1) (2018). https://doi.org/10.1080/15391523.2017.1401498
4. Al-Zawqari, A., Peumans, D., Vandersteen, G.: A flexible feature selection approach for predicting students' academic performance in online courses. Comput. Educ.: Artif. Intelli. **3** (2022). https://doi.org/10.1016/j.caeai.2022.100103

5. Alharbi, A.A., Allohibi, J.: A new hybrid classification algorithm for predicting student performance. AIMS Math. **9**(7) (2024). https://doi.org/10.3934/math.2024893
6. Alvarez-Garcia, M., Arenas-Parra, M., Ibar-Alonso, R.: Uncovering student profiles. an explainable cluster analysis approach to pisa 2022. Comput. Educ. **223** (2024). https://doi.org/10.1016/j.compedu.2024.105166
7. Bolaños-Rodríguez, E., Flores-Amador, C., López-Chau, A., Solis-Galindo, A.E., Zárate-Rosas, A.: Predicting academic performance in university students using machine learning. J. Eng. Technol. Indust. Appl. **10**(49) (2024). https://doi.org/10.5935/jetia.v10i49.1308
8. Bujang, S.D.A., et al.: Imbalanced classification methods for student grade prediction: a systematic literature review. IEEE Access **11** (2023). https://doi.org/10.1109/ACCESS.2022.3225404
9. Carhuamaca, V.M., Huamán, C.Q., Cangahuala, G.M.R., Muriel, P.F., Caso, J.G.: Predictive model to reduce undergraduate student dropout at the army scientific and technological institute of peru. In: Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology (2024). https://doi.org/10.18687/LACCEI2024.1.1.571
10. Delahoz-Domínguez, E.J., Hijón-Neira, R.: Recommender system for university degree selection: a socioeconomic and standardised test data approach. Applied Sciences (Switzerland) **14**(18) (2024). https://doi.org/10.3390/app14188311
11. Dervenis, C., Kyriatzis, V., Stoufis, S., Fitsilis, P.: Predicting students' performance using machine learning algorithms. In: ACM International Conference Proceeding Series (2022). https://doi.org/10.1145/3564982.3564990
12. Forid, M.S., Hafez, M., Khan, W.: Student satisfaction and retention: Impact of service quality and digital transformation. Market. Manage. Innov. **13**(4) (2022). https://doi.org/10.21272/mmi.2022.4-14
13. Guanin-Fajardo, J.H., Guaña-Moya, J., Casillas, J.: Predicting academic success of college students using machine learning techniques. Data **9**(4) (2024). https://doi.org/10.3390/data9040060
14. Hemdanou, A.L., Sefian, M.L., Achtoun, Y., Tahiri, I.: Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. Computers and Education: Artif. Intell. **7** (2024). https://doi.org/10.1016/j.caeai.2024.100301
15. Kaensar, C., Wongnin, W.: Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning. Eurasia J. Math., Sci. Technol. Educ. **19**(12) (2023). https://doi.org/10.29333/ejmste/13863
16. Kustitskaya, T.A., Esin, R.V., Vainshtein, Y.V., Noskov, M.V.: Hybrid approach to predicting learning success based on digital educational history for timely identification of at-risk students. Educ. Sci. **14**(6) (2024). https://doi.org/10.3390/educsci14060657
17. Lakhdar, Y., EL-Bendadi, K., Bakkas, B.: A new hybrid model to predict the performance of trainee teachers based on clustering and classification. J. Comput. Sci. **20**(9) (2024). https://doi.org/10.3844/jcssp.2024.1020.1029
18. Maniyan, S., Ghousi, R., Haeri, A.: Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. Comput. Educ.: Artif. Intell. **6** (2024). https://doi.org/10.1016/j.caeai.2024.100242
19. Muhammad, Y., Hassan, M.A., Almotairi, S., Farooq, K., Granelli, F., Ľubomíra Strážovská: The role of socioeconomic factors in improving the performance of

students based on intelligent computational approaches. Electronics (Switzerland) **12**(9) (2023). https://doi.org/10.3390/electronics12091982
20. Nachouki, M., Naaj, M.A.: Predicting student performance to improve academic advising using the random forest algorithm. Int. J. Distance Educ. Technol. **20**(1) (2022). https://doi.org/10.4018/IJDET.296702
21. Nafuri, A.F.M., Sani, N.S., Zainudin, N.F.A., Rahman, A.H.A., Aliff, M.: Clustering analysis for classifying student academic performance in higher education. Appl. Sci. (Switzerland) **12**(19) (2022). https://doi.org/10.3390/app12199467
22. Nakhipova, V., Bulbul, H.I., Kerimbekov, Y., Suleimenova, L., Umarova, Z., Adylbekova, E.: Integration of collaborative filtering into naive bayes method to enhance student performance prediction. Int. J. Inform. Commun. Technol. Educ. **20**(1) (2024). https://doi.org/10.4018/IJICTE.352512
23. Niu, K., Jia, B., Zhou, Y., Lu, G.: A hybrid model for predicting academic performance of engineering undergraduates. Int. J. Model., Simul. Sci. Comput. **14**(2) (2023). https://doi.org/10.1142/S1793962323500307
24. Poudyal, S., Mohammadi-Aragh, M.J., Ball, J.E.: Prediction of student academic performance using a hybrid 2d CNN model. Electronics (Switzerland) **11**(7) (2022). https://doi.org/10.3390/electronics11071005
25. Realinho, V., Machado, J., Baptista, L., Martins, M.V.: Predicting student dropout and academic success. Data **7**(11) (2022). https://doi.org/10.3390/data7110146
26. Roslan, N., Jamil, J.M., Shaharanee, I.N.M., Alawi, S.J.S.: Prediction of student dropout in malaysian's private higher education institute using data mining application. J. Adv. Res. Appl. Sci. Eng. Technol. **45**(2) (2025). https://doi.org/10.37934/araset.45.2.168176
27. Sánchez-Sánchez, A.M., Mello-Román, J.D., Segura, M., Hernández, A.: Identifying the determinants of academic success: A machine learning approach in Spanish higher education. Systems **12**(10) (2024). https://doi.org/10.3390/systems12100425
28. Valero, J.E., Navarro, Á.F., Larios, A.C.: Deserción universitaria: Evaluación de diferentes algoritmos de machine learning para su predicción. Como citar APA **XXVIII**, 362–375 (2022)
29. Villareal-Torres, H., Angeles-Morales, J., Marin-Rodrigez, W., Cano-Mejia, J.: Modelo de clasificación para la deserción estudiantil en las universidades públicas del perú. Revista de Ciencias Sociales **30**, 452–469 (2024)
30. Viveka, M., Priya, N.S.: An efficient hybrid deep learning framework for predicting student academic performance. Salud, Ciencia y Tecnologia - Serie de Conferencias **3** (2024). https://doi.org/10.56294/sctconf2024759
31. Vora, D.R., Rajamani, K.: A hybrid classification model for prediction of academic performance of students: a big data application. Evol. Intell. **15**(2) (2022). https://doi.org/10.1007/s12065-019-00303-9
32. Wu, J.: Integration of higher education student management and pedagogical concepts based on data-based decision making. Appl. Math. Nonlinear Sci. **9**(1) (2024). https://doi.org/10.2478/amns-2024-0991
33. Xie, Y.: Efficiency of hybrid decision tree algorithms in evaluating the academic performance of students. Int. J. Adv. Comput. Sci. Appl. **15**(2) (2024)
34. Yin, M., Cao, H., Yu, Z., Pan, X.: Manual label and machine learning in clustering and predicting student performance: a practice based on web-interactive teaching systems. Int. J. Web-Based Learn. Teach. Technolo. **19**(1) (2024). https://doi.org/10.4018/IJWLTT.347661

35. Yucebas, S.C., Yalpir, S., Genc, L., Dogan, M.: Price prediction and determination of the affecting variables of the real estate by using x-means clustering and cart decision trees. J. Univ. Comput. Sci. **30**(4) (2024). https://doi.org/10.3897/jucs.98733
36. Zhang, X., Yue, J.: Predictive modeling of student performance through classification with gaussian process models. Int. J. Adv. Comput. Sci. Appl. **15**(6) (2024)

Predictive Airfare Analysis Based on Machine Learning Models

Yasumy Maricely Jallo-Paccaya(✉), Jadyra Ch'aska Choque-Quispe, Thalia Quispe-Mamani, Rebeca Araceli Ccansaya-Soncco, Edwin Alvarez-Mamani, and Harley Vera-Olivera

Universidad Nacional de San Antonio Abad del Cusco, Cusco 08000, Peru
{204799,204795,204794,171879,edwin.alvarez,harley.vera}@unsaac.edu.pe

Abstract. This study addresses the growing need to predict airfare prices in highly variable contexts such as Perú, where `LATAM` Airlines stands out as the leading national carrier. In response to the complexity of dynamic pricing strategies, a comprehensive predictive solution is proposed with four main objectives: predicting airfares, estimating the optimal advance purchase period, identifying the most favorable time range to search for flights, and jointly predicting these variables through a multioutput model. To achieve this, a complete pipeline was designed, including automated data collection through web scraping using `Selenium` on LATAM's website, followed by preprocessing steps involving data cleaning, duplicate removal, outlier handling, and categorical variable encoding using both *One-hot encoding* and *Embedding-based encoding*. In the modeling stage, classical algorithms such as `XGBoost`, `Random Forest`, `Decision Tree`, and `K-Nearest Neighbors` (`KNN`) were implemented, alongside deep learning models such as `LSTM`, `GRU`, `TCN`, and `Transformers`. Hyperparameter optimization was carried out using the `Optuna` library, aiming to maximize performance in airfare prediction.

Keywords: airfare prediction · multioutput forecasting · One-hot encoding · Embedding-based encoding · machine learning models

1 Introduction

The prediction of airfares has become an increasingly relevant topic due to the complexity of dynamic pricing strategies implemented by airlines. In Peru, LATAM Airlines has the widest coverage of domestic destinations and leads the market with more than 60% market share, making it a representative case for the development of predictive models [1]. Anticipating its price behavior allows for optimizing purchasing decisions and designing user support tools.

Despite advances in predictive modeling techniques, there is still a lack of accessible and accurate systems capable of reliably estimating the main variables involved in efficient airline ticket purchasing. This study addresses four main objectives: 1) the prediction of airfares; 2) the estimation of the optimal advance

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 450–467, 2026.
https://doi.org/10.1007/978-3-032-20322-9_31

purchase period; 3) the identification of the most favorable time range to search; and 4) the joint prediction of these three variables using a multioutput model.

To achieve these objectives, a complete pipeline was designed, including automated data collection through web scraping using `Selenium` on the LATAM Airlines website, as well as preprocessing steps involving data cleaning, removal of duplicates, outlier handling, and encoding of categorical variables through both *One-hot encoding* and *Embedding-based encoding*. In the modeling stage, multiple approaches were explored to address the predictive objectives defined in this study. Both classical algorithms such as `eXtreme Gradient Boosting` (`XGBoost`), `Random Forest`, `Decision Tree`, and `K-Nearest Neighbors` (`KNN`), and deep learning models such as `Long Short-Term Memory` (`LSTM`), `Gated Re- current Unit` (`GRU`), `Temporal Convolutional Network` (`TCN`), and `Transfor- mers` were applied. These models were optimized using the `Optuna` library for efficient hyperparameter tuning and performance enhancement in predicting the target variables related to airfares.

The performance evaluation was carried out using standard predictive metrics, including *Mean Absolute Error* (*MAE*), *Mean Squared Error* (*MSE*), *Root Mean Squared Error* (*RMSE*), *R-squared* (R^2), *Accuracy*, *F1-score*, and execution time.

Overall, this study offers an integrated solution that combines automation, advanced preprocessing techniques, and diverse modeling approaches to address a real world problem with potential impact on both users and airline operators. Its replicable, scalable, and context-aware nature in the Peruvian market positions it as a relevant contribution to the development of intelligent recommendation systems in highly variable environments.

This paper is organized as follows: the related work section presents the context and prior studies; the methodology section describes the proposed approach and techniques used; the experiments section outlines the experimental design and procedures followed; the results section reports the main quantitative findings; the discussion section analyzes and interprets these findings; and finally, the conclusions section summarizes the study's contributions and outlines potential future research directions.

2 Related Work

Various studies have addressed airline ticket price prediction using machine learning techniques, employing approaches ranging from classical regression models to broader predictive analytics platforms. Three thematic clusters were identified according to their main focus: classical regression and decision tree models, advanced approaches such as deep or quantum learning, and ranking strategies for purchase decision support.

A first group of studies applied classical models such as `Linear Regression`, `Decision Tree`, `Random Forest`, `Support Vector Machine`, and `XGBoost`. Rajankar et al. [2], Krishna et al. [3], and Mukane et al. [4] worked with local

datasets and reported that `Random Forest` outperformed simpler models. Fadhil et al. [5] reached a similar conclusion with an accuracy of 87.4%. Joshitta et al. [6] integrated five models into a predictive system, concluding that `Random Forest` with feature selection achieved the best performance (RMSE = 1.98, $R^2 = 0.81$). Meanwhile, Vaishnavi et al. [7] and Iswarya [8] identified `XGBoost` as the most accurate model, achieving values up to 90.17%. Korkmaz [9] proposed `Gaussian Process Regression (GPR)` as an alternative, obtaining the best result with $R^2 = 0.90$ in tests.

A second set of studies incorporates more sophisticated approaches such as deep learning and quantum learning. Kalampokas et al. [10] employed convolutional neural networks (CNNs) and quantum models such as `Quantum MLP` and `Quantum SVM`, achieving accuracies between 89% and 99% on a dataset of over 136,000 flights. Although the results are promising, these models present higher complexity and have not yet been implemented in production.

A third approach, adopted by Oglakcioglu et al. [11], is based on *Learning to Rank* techniques such as `Ranking SVM` and `LambdaMART`. Unlike the previous ones, these methods do not predict prices but rather prioritize optimal purchase dates, generating strategic recommendations for users.

In summary, the studies converge on the effectiveness of models such as `Random Forest`, `XGBoost`, and `GPR` for price prediction. Advanced deep and quantum learning models show higher accuracy but present technical challenges.

In this context, Abdella et al. [12] conducted a comprehensive review of existing methods for predicting airline ticket prices and demand, classifying approaches into two main categories: customer-side models, focused on determining the optimal time to buy, and airline-side models, aimed at estimating demand and optimizing revenue. The authors highlight that most studies use internal variables such as price history and flight dates, while external factors like online searches, social media, weather conditions, or events are rarely considered. They also recommend the use of deep learning techniques and vector representations through *embeddings* to improve prediction accuracy. This work serves as a solid theoretical foundation for future research seeking to integrate multiple data sources and more advanced AI approaches in airline price prediction.

3 Methodology

The primary objective of this research is to predict four target variables in the context of airline ticket purchases: airfare, days in advance, the most convenient time range, and a joint prediction of these variables. To achieve this, a methodological pipeline was designed that integrates data collection techniques, categorical variable processing, the training of classical machine learning models and deep neural networks, as well as comparative evaluation of results.

Figure 1 presents the proposed methodological pipeline for predicting airfares, advance purchase days, optimal time ranges for booking, and the joint prediction of these variables. The process begins with the automated collection

of data through web scraping techniques, which are then stored in a structured dataframe. Next, a preprocessing and data transformation phase is carried out, including approaches such as *One-hot encoding* and *Embedding-based encoding*, aimed at appropriately representing categorical variables. These transformations allow for the creation or modification of relevant features in order to enhance predictive performance.

Once the preprocessed dataset is ready, the experimentation phase is conducted using four classical machine learning models (`XGBoost, KNN, Decision Tree, and Random Forest`) and four deep learning models (`LSTM, GRU, TCN, and Transformers`). This stage includes the selection of target variables, hyperparameter optimization using tools such as `Optuna`, and the training of models. Finally, the results are evaluated and compared through statistical analysis, enabling the validation of model performance and the assessment of their behavior across the different prediction objectives.

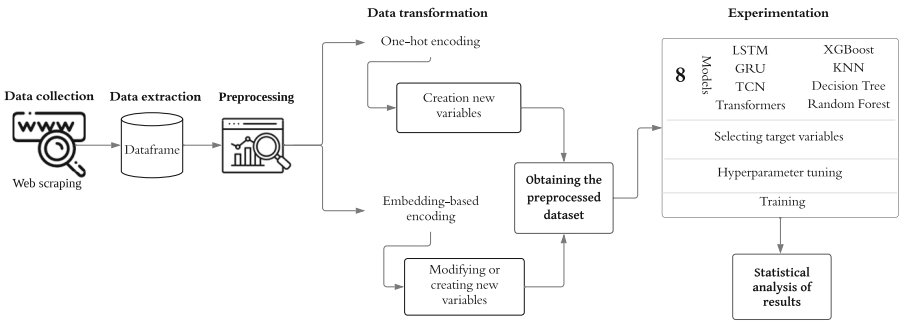


Fig. 1. Diagram of the methodological pipeline for airfare prediction.

3.1 Dataset

The dataset used in this study was collected from the official LATAM Airlines website through automated web scraping techniques applied to domestic flights within Peru. For each combination of origin, destination, and departure date, relevant features were extracted, including origin and destination cities, departure and arrival times, flight duration, fare (USD), operating airline, flight type (direct or connecting), departure date, and query date.

This dataset reflects the dynamic nature of the national airline market, where prices fluctuate depending on purchase anticipation, time range, and other contextual factors. The final structure combines categorical and temporal variables, which were later transformed for predictive analysis.

Tools Used: To automate navigation and data extraction, the `Python` programming language was used, along with specialized libraries such as `Selenium`,

`Schedule`, and `Time`. These tools allowed for simulating real searches, controlling execution intervals, and performing periodic web scraping tasks. Interaction with dynamic pages was made possible through the use of Chrome WebDriver, which emulates a real browser and facilitates the capture of dynamically generated content.

Data Collection: The data collection process was conducted through an automated system that queried various combinations of origin, destination, and departure dates directly from the official LATAM Airlines website. The data were extracted using *XPath* selectors, allowing precise identification of key webpage elements.

The collection was performed continuously between October 11, 2024, and March 3, 2025, with a five-minute interval between each query. To prevent overloading the website's servers, deliberate pauses were introduced between requests, in accordance with ethical standards and best practices. In addition, error-handling mechanisms were implemented to ensure system stability in the event of interruptions or content-loading failures.

The process was carried out responsibly and transparently, relying exclusively on publicly available fare information, without accessing any personal, confidential, or transactional data. This approach is in line with the ethical principles of the project, ensuring compliance with the website's policies and the legitimate use of the data.

Although all variables were initially stored as *object-type* data, they were subsequently converted to appropriate data types according to their content. Finally, the collected data were stored in CSV files, following the structure described in Table 1.

Table 1. Structure of the collected dataset.

Variable	Data Type	Description
Origin	category	City of flight origin
Departure time	datetime	Scheduled departure time
Arrival time	datetime	Estimated arrival time
Destination	category	City of flight destination
Duration	datetime	Total flight time
Fare	float	Ticket price in US dollars (USD)
Operated by	string	Airline operating the flight
Direct	string	Indicator of direct or connecting flight
Departure date	datetime	Scheduled departure date
Query date	datetime	Date the search was performed

3.2 Data Preprocessing

Data preprocessing is a fundamental stage in building predictive models, as it enables proper preparation of the information before training. In this study, two approaches were implemented for handling categorical variables: *One-hot encoding* and *Embedding-based encoding*. The procedures applied for both encodings, as well as their specific characteristics, are described below.

Common Preprocessing for both Approaches: During this stage, general procedures were carried out to ensure data quality and consistency prior to training, regardless of the encoding strategy applied. First, the data were loaded from CSV files and an initial exploration was conducted to identify their structure, variable types, and possible errors. Duplicate records were then removed using the `duplicated()` function from the `Pandas` library. For handling missing values, a threshold of 5% per variable was established; since all attributes had only one missing value (less than 0.01%), they were imputed using the mode. Finally, the interquartile range (IQR) method was applied to detect and correct outliers in key variables such as flight duration and fare. The boundaries were calculated using the following formulas:

$$\text{Upper bound} = Q_3 + 1.5 \times IQR \quad \text{and} \quad \text{Lower bound} = Q_1 - 1.5 \times IQR \quad (1)$$

As shown in Eq. 1, these bounds made it possible to adjust out-of-range values without removing relevant records. Next, temporal variables were transformed to facilitate their use in the models: flight duration was converted to minutes, schedules were grouped into two hour intervals, and the variable *days_until_flight* was created. The duration format was standardised, and the flight-type variable was binarised (1 for direct flights, 0 for connecting flights). Finally, a correlation analysis using the *Pearson* coefficient revealed relevant relationships, such as the negative correlation between fare and advance-purchase days (–0.1). Constant variables such as the operating airline and the departure date were removed to reduce dimensionality, as they provided no predictive value.

Preprocessing with *One-Hot Encoding*: The *One-hot encoding* approach was used to transform the categorical variables indicating the origin and destination cities of flights into binary vectors. Although these variables have moderately high cardinality due to the number of cities considered, this method was chosen for its ability to represent categories without imposing an artificial order.

This approach prevents the model from assuming nonexistent ordinal relationships and is compatible with classical models such as `Decision Tree` and `XGBoost`, facilitating efficient representation of categorical variables. Furthermore, assigning one column per city improves interpretability and importance analysis. Although it increases the dimensionality of the dataset, the limited number of categories kept the computational cost manageable.

Preprocessing with *Embedding-Based Encoding*: For categorical variables such as time ranges (for query, departure, and arrival), as well as origin and destination cities, *Embedding-based encoding* was applied due to its ability to capture

latent relationships between categories within a lower-dimensional vector space. This approach is particularly useful in fare prediction, where the time of day or specific route combinations can significantly influence pricing. In this study, *Embedding-based encoding* were applied to time ranges grouped into 12 two-hour intervals and to the relevant cities, enabling the model to identify meaningful temporal and spatial patterns. This improved the generalization capability of the model compared to representations such as *One-hot encoding*.

Moreover, the compact structure of *Embedding-based encoding* representations is especially well suited for neural network-based models, which benefit from implicit relationships among categories when modeling complex dynamics.

Finally, the data preprocessed using *One-hot encoding* and *Embedding-based encoding* were stored in CSV files, ensuring traceability and facilitating comparisons and future adjustments without compromising dataset integrity. This is essential for developing robust and accurate predictive models for airfare estimation.

3.3 Model Description

To address the prediction of airline fares considering temporal, categorical, and numerical variables, eight representative models were selected, encompassing both classical machine learning methods and deep neural network architectures. This combination enables the evaluation of the effectiveness of traditional methods against sequence-specialized architectures.

For the neural network–based models, `LSTM` and `GRU` were considered, both capable of capturing long-term dependencies, with `GRU` being a lighter variant. `TCN` employs causal convolutions that capture temporal patterns in parallel, enhancing training efficiency. The `Transformer`, based on attention mechanisms, identifies complex relationships without processing inputs sequentially.

Regarding classical models, `Decision Tree` was included for its interpretability and speed; `Random Forest` and `XGBoost` stand out for their precision and generalization through ensemble techniques; and `KNN`, a non-parametric proximity-based method that requires no formal training. Together, these models allow assessing various levels of complexity, accuracy, and computational efficiency.

To ensure a fair and reproducible evaluation, all models were trained under a unified experimental framework, with hyperparameter optimization performed using `Optuna` via Bayesian search. Each model was trained over ten iterations of 100 epochs, balancing exploration and computational efficiency.

3.4 Description of the Evaluation Metrics Used by Objectives

This study addressed four objectives: the first was to predict the flight fare; the second, to estimate the optimal number of days in advance to purchase the ticket; and the third, to identify the most convenient time range to query and acquire the ticket. The fourth objective integrated the previous three through a joint prediction approach using multioutput models, enabling the evaluation of the system's ability to solve mixed regression and classification tasks simultaneously.

For Objectives 1 and 2, which involve continuous variables, regression metrics were used: R^2, *MAE*, *MSE*, and *RMSE*. For objective 3, a classification task, metrics such as *accuracy*, *recall*, and *F1-score* were employed, which are particularly useful for evaluating performance in the presence of class imbalance. Regarding objective 4, the metrics applied corresponded to the nature of each output: R^2, *MAE*, *MSE*, and *RMSE* for the continuous variables (fare and days in advance), and *accuracy* for the categorical variable (time range), ensuring a coherent and specific evaluation for each component of the multioutput model.

4 Experiments

The experiments were conducted in a high-performance computing environment equipped with 128 GB of DDR4 ECC RAM, *Intel Xeon* processors, and support for parallel processing via *NVIDIA Quadro GP100* GPU. Table 2 summarizes the computational resource usage by model during experimentation.

Table 2. Summary of computational resource usage by model during experimentation.

Model	Model Type	CPU	GPU
`LSTM`	Recurrent Neural Network	✓	✓
`GRU`	Recurrent Neural Network	✓	✓
`TCN`	Temporal Convolutional Network	✓	✓
`Transformers`	Sequential Transformer	✓	✓
`Random Forest`	Ensemble Model	✓	
`Decision Tree`	Traditional Model	✓	
`KNN`	Traditional Model	✓	
`XGBoost`	Ensemble Model	✓	

As part of the experimental process, the *Optuna* library was employed for automatic hyperparameter optimization, adjusting each model according to the encoding strategy used (*One-hot encoding* or *Embedding-based encoding*) and the corresponding prediction objective. Through Bayesian optimization and early-pruning strategies, configurations were identified that maximized model performance with respect to the target metric.

The results indicated that both the number and type of optimal hyperparameters varied across model architectures and encoding schemes, reflecting the influence of model structure on predictive performance. Table 3 presents a summary of the optimal values obtained exclusively for Objective 1, while the complete table containing results for all objectives is available in the public repository: https://github.com/yib-thv/latam-price-prediction/

To ensure the reproducibility of the study, all source code including training, evaluation, preprocessing, and optimization scripts has been documented and organized within the same repository.

Table 3. Optimal hyperparameters for Objective 1

Model	One-hot encoding	Embedding-based encoding
`LSTM`	lstm units: 76; dense units: 109; dropout rate: 0.2040; learning rate: 0.0058	embedding dim: 4; lstm units: 64; dense units: 32; dropout rate: 0.4377; learning rate: 0.0063
`GRU`	gru units: 115; dense units: 118; dropout rate: 0.1946; learning rate: 0.0004	embedding dim: 15; gru units: 113; dense units: 120; learning rate: 0.0011
`TCN`	tcn filters: 38; kernel size: 2; dense units: 114; dropout rate: 0.1891; learning rate: 0.0011	embedding dim: 13; tcn filters: 78; dense units: 108; learning rate: 0.0002
`Transformer`	num heads: 3; key dim: 47; ff dim: 69; dense units: 126; dropout rate: 0.2321; learning rate: 0.0022	embedding dim: 4; num heads: 6; ff dim: 112; dropout rate: 0.4729; learning rate: 0.0094
`Random Forest`	n estimators: 400; max depth: 30; min samples split: 9; min samples leaf: 2	n estimators: 61; max depth: 18; min samples split: 10; min samples leaf: 2; max features: log2
`Decision Tree`	max depth: 28; min samples split: 7; min samples leaf: 2	max depth: 19; min samples split: 10; min samples leaf: 4; max features: log2
`KNN`	n neighbors: 6; weights: uniform; p: 1	n neighbors: 15; weights: distance; p: 2
`XGBoost`	n estimators: 199; max depth: 8; subsample: 0.8406; colsample bytree: 0.9596	n estimators: 64; max depth: 8; learning rate: 0.2101; subsample: 0.8308; colsample bytree: 0.6683; gamma: 3.8772; reg alpha: 0.3777; reg lambda: 0.0909

5 Results

This section presents the outcomes for the four prediction objectives, comparing the performance of classical models and neural networks.

Each model was trained individually for each objective, except for Objective 4, where joint predictions were performed. Performance was evaluated using metrics appropriate to the nature of each variable (regression or classification).

5.1 One-Hot Encoding

The following results were obtained using *One-hot encoding*. Tables 4, 5, 6, and 7 present the performance of the eight evaluated models, organized according to the defined prediction objectives.

Objective 1: Table 4 summarizes the results for fare prediction. The `Random Forest` model achieved the lowest *MAE* of 4.90 and a high R^2 of 0.96, combining accuracy with low computational cost. `Decision Tree` and `XGBoost` also showed competitive results, while `GRU` and `LSTM` performed well but required longer training times. `Transformers` obtained the lowest performance, with an R^2 of 0.82.

Table 4. Model results for air fare prediction.

Model	MAE	MSE	RMSE	R^2	Time (seconds)
`LSTM`	13.62	415.69	20.39	0.85	17059.06
`GRU`	12.75	394.88	19.87	0.85	16763.93
`TCN`	12.61	399.95	20.00	0.85	88563.02
`Transformers`	14.24	497.49	22.30	**0.82**	27091.27
`Random Forest`	**4.90**	110.13	10.49	**0.96**	58.81
`Decision Tree`	**4.90**	141.15	11.88	**0.95**	5.21
`KNN`	10.01	334.71	18.30	0.88	55.29
`XGBoost`	**6.28**	117.44	10.84	**0.96**	3.71

Objective 2: Table 5 shows that `Random Forest` achieved the best results, combining high accuracy *MAE* of 8.22, R^2 of 0.88 with low execution time. `XGBoost` was similarly accurate but slower, while `Transformers` and `Decision Tree` offered acceptable performance quickly. `TCN`, `LSTM`, and `KNN` showed the weakest results.

Table 5. Results of *One-hot encoding* models for advance days prediction.

Model	MAE	MSE	RMSE	R^2	Time (seconds)
`LSTM`	78.16	7292.73	85.39	0.21	354.24
`GRU`	10.62	350.95	18.73	0.25	51636.87
`TCN`	14.39	516.62	22.73	0.07	7568.29
`Transformers`	14.92	848.62	29.13	0.50	4.63
`Random Forest`	**8.22**	**207.26**	**14.39**	**0.88**	**25.05**
`Decision Tree`	15.01	835.17	28.89	0.51	19.53
`KNN`	24.02	1649.65	40.62	0.03	21.36
`XGBoost`	8.23	248.93	15.78	0.47	824.42

Objective 3: Table 6 presents the results for flight time range prediction, where the `Decision Tree` model achieved the best performance with an *F1*-score of 0.64 and balanced precision and recall, followed by `Random Forest` and `KNN` with notably lower values. Among deep learning approaches, `Transformers` obtained the highest *F1*-score of 0.21, while `LSTM` and `XGBoost` showed the weakest results. Regarding efficiency, `Decision Tree` and `KNN` had short training times, whereas neural networks required considerably longer execution times without reaching competitive accuracy.

Table 6. Results of *One-hot encoding* models for flight time range prediction.

Model	Accuracy	Precision	Recall	F1-score	Time (seconds)
`LSTM`	0.15	0.12	0.15	0.09	2059.12
`GRU`	0.16	0.13	0.19	0.16	1447.18
`TCN`	0.18	0.15	0.18	0.15	577.45
`Transformers`	0.22	0.24	0.22	**0.21**	597.94
`Random Forest`	**0.35**	0.35	0.35	**0.34**	140.47
`Decision Tree`	0.64	0.67	0.63	**0.64**	15.94
`KNN`	0.24	0.25	0.24	**0.24**	49.32
`XGBoost`	0.14	0.10	0.14	0.09	26.12

Objective 4: Table 7 presents the results for the simultaneous prediction of fare, advance purchase days, and time range. `Random Forest` achieved the best overall performance, with the lowest errors in fare prediction (MAE of 16.71, R^2 of 0.72) and advance days estimation (MAE of 24.89, R^2 of 0.39), while requiring only 9.79 s of training.

`KNN` obtained competitive results in fare and advance days prediction, and it achieved the highest performance in time range classification (accuracy of 0.12, *F1-score* of 0.07). In terms of efficiency, `Decision Tree` was the fastest model (0.89 s), while deep learning models such as `TCN` and `Transformers` required significantly longer training times with lower predictive accuracy.

Table 7. Multioutput model results for the three prediction objectives (Objective 4).

Model	Fare		Advance Days		Time Range		Time (seconds)
	MAE	R^2	MAE	R^2	Accuracy	F1-score	
`LSTM`	25.68	0.51	30.01	0.14	0.12	0.06	1199.44
`GRU`	24.82	0.52	30.01	0.15	0.12	0.06	1393.37
`TCN`	26.60	0.49	31.37	0.13	0.12	0.06	1819.95
`Transformers`	26.21	0.50	29.98	0.14	0.12	0.06	2321.31
`Random Forest`	**16.71**	**0.72**	**24.89**	**0.39**	0.12	0.06	**9.79**
`Decision Tree`	19.61	0.63	27.09	0.29	0.12	0.06	**0.89**
`KNN`	16.89	0.70	25.01	0.35	**0.12**	**0.07**	16.28
`XGBoost`	20.25	0.64	27.59	0.30	0.12	0.06	2.43

5.2 Embedding-Based Encoding

The following are the results obtained using *Embedding-based encoding*. Tables 8, 9, 10, and 11 present the performance of the eight evaluated models, organized by objective.

Objective 1: Table 8 shows that the `KNN` model achieved the best performance, with a *MAE* of 16.88 and an R^2 of 0.70. The lowest R^2 was obtained by the `LSTM` model, with a value of 0.26, which also exhibited the highest *MAE* of 34.86. Regarding execution time, the fastest model was `Decision Tree`, requiring 0.18 s.

Table 8. Model results using *Embedding-based encoding* for airfare prediction (Objective 1).

Model	MAE	MSE	RMSE	R^2	Time (seconds)
`LSTM`	**34.86**	2598.10	50.97	**0.26**	381.24
`GRU`	30.67	2108.84	45.92	0.40	451.57
`TCN`	22.60	1063.21	32.61	0.70	2993.13
`Transformers`	32.65	1984.97	44.55	0.44	274.74
`Random Forest`	19.50	959.98	30.13	0.67	8.12
`Decision Tree`	19.52	1288.07	31.46	0.64	**0.18**
`KNN`	**16.88**	813.63	28.52	**0.70**	0.46
`XGBoost`	19.82	917.52	30.29	0.66	0.26

Objective 2: Table 9 shows that the `GRU` and `Transformers` models achieved the best performance in predicting optimal advance days, with an R^2 of 0.66 and low *MAE* values. In contrast, classical models like `Random Forest` and `Decision Tree` performed worse, with R^2 below 0.20. Although deep learning models require more training time, their higher accuracy justifies their use for complex temporal data.

Objective 3: Table 10 indicates that traditional models such as `GRU` and `Decision Tree` stood out significantly, achieving the best overall metrics, particularly in terms of *F1-score*, with values of 0.54 and 0.61, respectively. Additionally, the `Transformers` model maintained competitive performance, with an *F1-score* of 0.51.

Table 9. Model performance for optimal days in advance prediction.

Model	MAE	MSE	RMSE	R^2	Time (seconds)
LSTM	18.37	653.10	25.56	0.62	19318.21
GRU	**17.18**	590.54	24.30	**0.66**	1866.09
TCN	17.11	602.63	24.55	0.65	1803.25
Transformers	16.79	579.84	24.08	**0.66**	3917.90
Random Forest	29.76	1387.75	37.25	0.19	81.82
Decision Tree	29.79	1406.69	37.51	0.18	0.65
KNN	24.89	1075.85	32.80	0.37	19.61
XGBoost	29.79	1406.56	37.50	0.18	6.23

Table 10. Model performance for flight time range prediction.

Model	Accuracy	Precision	Recall	F1-score	Time (seconds)
LSTM	0.1986	0.1864	0.1986	0.1686	20817.06
GRU	**0.54**	**0.56**	**0.54**	**0.54**	19299.48
TCN	0.2569	0.2549	0.2569	0.2422	17598.86
Transformers	0.51	0.52	0.5071	0.51	26805.73
Random Forest	0.21	0.34	0.51	0.41	133.4751
Decision Tree	**0.66**	**0.47**	**0.58**	**0.61**	0.56
KNN	0.2441	0.2458	0.2441	0.2401	49.32
XGBoost	0.14	0.25	0.57	0.35	26.16

Objective 4: Table 11 demonstrates that the Decision Tree model excelled in regression, achieving the lowest *MAE* of 16.67 and the highest R^2 of 0.71 for fare prediction. KNN also exhibited competitive performance, with a *MAE* of 16.91 and a low execution time of 0.02 s. Random Forest attained an R^2 of 0.68, maintaining a balance between accuracy and efficiency. Meanwhile, TCN showed acceptable performance across the three outputs, with an *MAE* of 21.25 and an R^2 of 0.59 for fare prediction, but at the cost of significantly higher computational time. In contrast, XGBoost exhibited the poorest performance, with an R^2 of 0.20 and an *MAE* of 38.76, indicating limitations in its generalization capability under this joint prediction scheme.

5.3 Training Time *One-Hot Encoding* vs. *Embedding-Based Encoding*

As shown in Fig. 2, the results for Objective 1 show that models trained using *Embedding-based encoding* exhibit significantly higher training times compared to those using *One-hot encoding*. This difference is noticeable for models such as LSTM and TCN, with approximately 3.8 h and 2.4 h, respectively. In contrast,

Table 11. Multioutput model results for the three output targets.

Model	Fare		Advance Days		Time Range		*Time (seconds)*
	MAE	R^2	MAE	R^2	Accuracy	F1-score	
`LSTM`	22.41	0.58	29.16	0.19	0.17	0.13	1961.34
`GRU`	22.10	0.58	29.70	0.19	0.17	0.13	2176.67
`TCN`	**21.25**	**0.59**	29.48	**0.20**	0.15	0.09	16145.49
`Transformers`	23.72	0.53	29.62	0.15	0.17	0.12	2683.56
`Random Forest`	18.76	**0.68**	28.60	0.26	0.16	0.10	206.51
`Decision Tree`	**16.67**	**0.71**	24.99	0.37	0.16	0.11	3.29
`KNN`	**16.91**	0.70	25.03	0.35	0.13	0.12	*0.02*
`XGBoost`	**38.76**	**0.20**	33.19	0.08	0.15	0.10	1.71

models such as `Decision Tree`, `Random Forest`, and `KNN` show lower training times with both encoding strategies. This indicates that while *Embedding-based encoding* improves representational power, it comes at a high computational cost.

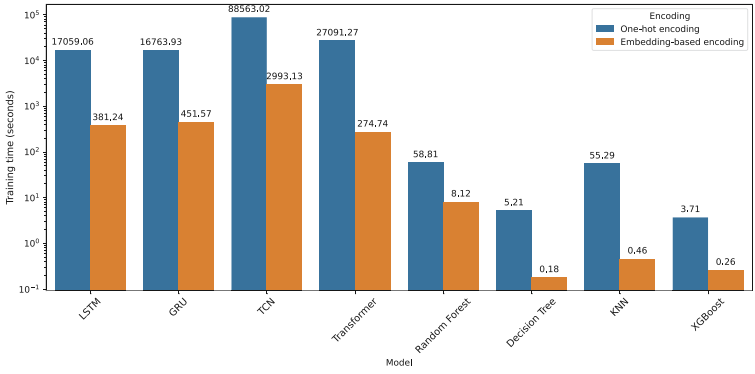


Fig. 2. Comparison of training time for price prediction.

As shown in Fig. 3, the results from Objective 2 reveal significant differences in training times between the two encoding approaches. Using *embedding-based encoding*, the `LSTM`, `GRU`, `TCN`, and `Transformer` models required 19,318, 1,866, 1,803, and 3,917 s, respectively, with durations ranging from approximately 30 min to over 5 h. In contrast, with *One-hot encoding*, the corresponding training times were 354, 51,636, 7,568, and 4.63 s. These findings highlight that in certain cases such as with `GRU` and `TCN` this approach may entail higher computational cost than the use of embeddings.

Figure 4 reveals notable differences in training times between the encoding strategies evaluated in Objective 3. With *Embedding-based encoding*, models such

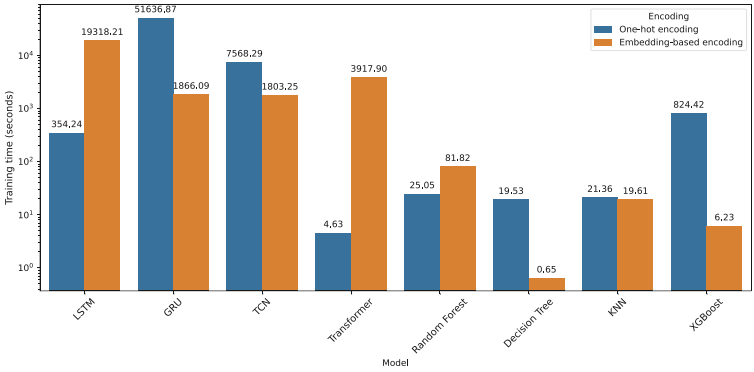


Fig. 3. Comparison of training time for *days-to-flight* prediction.

as `LSTM`, `GRU`, `TCN`, and `Transformer` require between 4 and 7 h of training. In contrast, under the *One-hot encoding* approach, most models are trained in under 30 min.

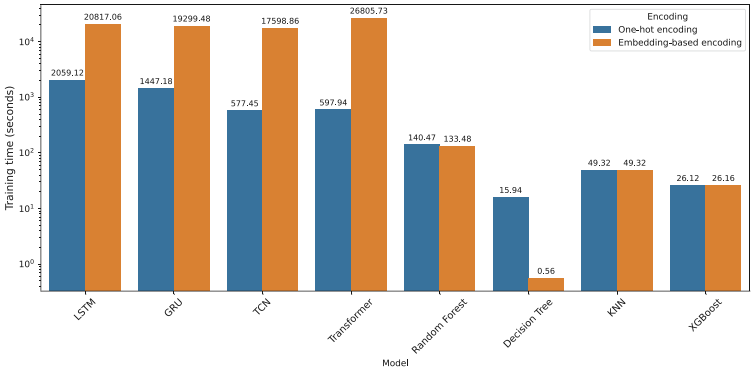


Fig. 4. Comparison of training time for time range prediction.

Figure 5 shows the results for Objective 4, where `TCN` using *Embedding-based encoding* reached the highest training time at over 1.6 h. Meanwhile, `LSTM` and `GRU` exceeded 9 min. With *One-hot encoding*, most models trained in under 2 min, except for deep neural networks, which took up to 38 min. These results highlight the substantial computational cost of *Embedding-based encoding* in multioutput architectures.

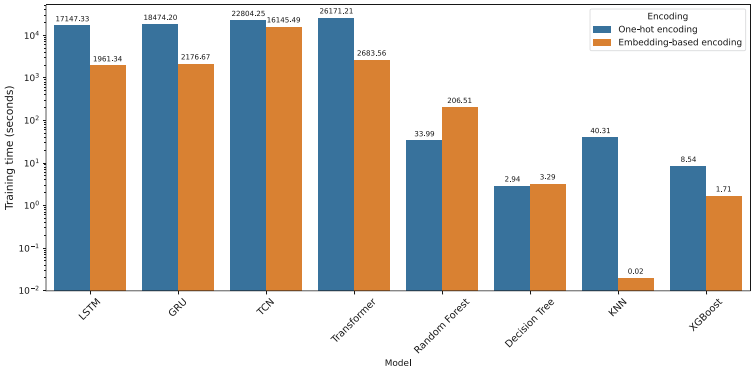


Fig. 5. Comparison of training time for a multi-output model.

6 Discussion

The results of this study reveal notable differences between classical models and deep neural networks across all prediction objectives, considering both *One-hot* and *Embedding*-based encoding.

With *One-hot encoding*, classical models such as `Decision Tree`, `KNN`, and `Random Forest` achieved competitive or superior performance in regression tasks, combining high accuracy with low computation times, as reported by Krishna et al. [3] and supported by Fadhil et al. [5].

With *Embedding*-based encoding, neural networks such as `TCN`, `GRU`, and `Transformers` improved their predictive capacity, particularly in Objectives 1, 2, and 4, but at the cost of significantly higher training times, as noted by Iswarya et al. [8] and Mukane et al. [4].

For classification tasks (Objective 3), classical models with *Embedding*-based encoding achieved the best metrics, with `KNN` offering the most balanced performance, as highlighted by Epra et al. [7]. Overall, classical models proved more computationally efficient, completing training in seconds compared to the hours required by deep neural networks.

7 Conclusions

This study presents an end-to-end pipeline for predicting the optimal time to purchase airline tickets, encompassing automated data collection, preprocessing, and experimentation with multiple machine learning and deep learning models. The evaluation was conducted using two encoding schemes, namely *One-hot encoding* and *Embedding-based encoding*, addressing four objectives: (1) predicting the published fare, (2) estimating the optimal purchase lead time, (3) identifying the most convenient time range, and (4) jointly predicting these variables through a *multi-output* model.

Results indicate that, with *One-hot encoding*, classical models such as `Random Forest` and `Decision Tree` achieved the best trade-off between accuracy and efficiency across most objectives. With *Embedding*-based encoding, `KNN` excelled in fare prediction, while `GRU` and `Transformers` outperformed classical models in estimating purchase lead time and time range classification.

Overall, classical models are more suitable when computational resources are limited, whereas deep learning models provide higher accuracy when greater computational capacity is available. Data collection from LATAM Airlines is still ongoing, which will strengthen the dataset and improve model robustness. Future work includes expanding coverage to other airlines to increase data diversity, applying advanced regularization and semi-supervised learning techniques, developing a real-time interactive recommendation system that provides personalized alerts to users, and validating the system in real-world scenarios with live data and dynamic market conditions to enhance its practical applicability.

References

1. Infraestructura Vial. Latam perú expande vuelos diarios a más destinos nacionales (2025). https://infraestructuravial.pe/noticias/latam-peru-expande-vuelos-diarios-a-mas-destinos-nacionales/. Available at: Infraestructura Vial
2. Rajankar, S., Sakharkar, N., Rajankar, O.: Predicción del precio de un billete de avión con el uso de algoritmos de aprendizaje automático. Int. J. Sci. Technol. Res. **8**, 3297–3300 (2019)
3. Krishna, M., Sharitha, G., Ganesh, M., Kumar, A., Karthika, G.: Flight ticket price prediction using regression models. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) **10**(3) (2022)
4. Mukane, J., Pawar, S., Pawar, S., Muley, G.: Predicción del precio de los billetes de avión mediante aprendizaje automático. Revista Internacional de Investigación en Ciencia Aplicada y Tecnología de Ingeniería (2022)
5. Fadhil, H.M., Abdullah, M.N., Younis, M.I.: A framework for predicting airfare prices using machine learning. Iraqi J. Comput. Commun. Control Syst. Eng. **22**(3), 81–88 (2022)
6. Joshitta, S.M., Sunil, M.P., Alfurhood, B.S., Bodhankar, A, Ch. Sreedevi, Khanna, R.: The integration of machine learning technique with the existing system to predict the flight prices. In: 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 398–402 (2023)
7. Vaishnavi, K.D.V.N., Hima Bindu, L., Satwika, M., Udaya Lakshmi, K., Harini, M., Ashok, N.: Flight fare prediction using machine learning. EPRA Int. J. Res. Dev. **8**(10), 245–250 (2023)
8. Iswarya, G.: Predicción de precios de billetes de avión mediante aprendizaje automático. Revista Internacional de Investigación Científica en Ingeniería y Gestión (2024)
9. Korkmaz, H.: Prediction of airline ticket price using machine learning method. J. Transp. Logist. (2024)
10. Kalampokas, T., Tziridis, K., Kalampokas, N., Nikolaou, A., Vrochidou, E., Papakostas, G.A.: A holistic approach on airfare price prediction using machine learning techniques. IEEE Access **11**, 46627–46643 (2023)

11. Oglakcioglu, K., Can, Y.S., Alagoz, F.: Prediction of optimal flight ticket purchase timing by using learning to rank methods. In: 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–6 (2023)
12. Abdella, J.A., Abdela, K., Getahun, F., Shah, M.: Airline ticket price and demand prediction: A survey. J. King Saud Univ. - Comput. Inf. Sci. **33**(4), 375–391 (2021)

Mobile App for Detecting Work-Related Stress Using Acoustic Biomarkers and Machine Learning

Claudia Lucia Rozas Gamero[1], Oscar Heredia Miranda[1], Eduardo Díaz[1(✉)], and Edgar Ceh-Varela[2]

[1] Universidad Peruana de Ciencias Aplicadas, Prolongación Primavera 2390, Lima 15023, Peru
{u201922910,u201913874,pcsijord}@upc.edu.pe
[2] Department of Mathematical Sciences, Eastern New Mexico University, Portales, NM, USA 88130
eduardo.ceh@enmu.edu

Abstract. The rising prevalence of work-related stress, particularly in remote settings, requires innovative and accessible monitoring solutions. In response, this paper details the development and validation of a mobile application that uses machine learning to detect work-related stress through acoustic biomarkers from the user's voice. The system relies on a Decision Tree model to analyze key vocal features, including Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and Root Mean Square (RMS) energy, for stress classification. Model robustness and generalizability were rigorously assessed using Leave-One-Out cross-validation. The final model achieved a 75% in accuracy, precision, and recall. These findings validate the efficacy of using acoustic biomarkers as non-invasive indicators of stress. This research contributes to a functional prototype, offering a significant step toward practical, low-cost tools for periodic mental health monitoring scheduled by leaders in the context of modern work environments.

Keywords: work stress · machine learning · employees · acoustic biomarkers

1 Introduction

Work-related stress has become an increasing concern due to the shift to remote work, especially after the pandemic. While this work model offers flexibility and autonomy, it has also brought risk factors like social isolation, work overload, and difficulty separating personal and professional life. Several studies have pointed out that remote work can increase the overall workload, exposing employees to job insecurity and, as a result, negatively impacting their engagement and emotional well-being [1]. At the same time, a specific type of stress, known as hindrance stress, has emerged. It not only affects individual performance but also family well-being, as workers find it hard to set boundaries between their work and personal lives [2]. The impact of telework has also been studied from an occupational health perspective. A study conducted in India found that work-related stress significantly affects motivation, job satisfaction, and performance,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 468–479, 2026.
https://doi.org/10.1007/978-3-032-20322-9_32

mainly due to the lack of co-worker interaction and prolonged isolation [3]. On top of that, the search for meaning at work can lead to emotional overload and burnout [4]. Even with the growing body of literature on remote work, there is still a gap when it comes to early detection of stress indicators in remote workers. In this context, using machine learning and acoustic biomarkers is emerging as an innovative way to identify stress patterns in human speech. This could allow for timely intervention before stress negatively affects employee performance or health [3, 5]. Recent studies have already shown the potential of wearable devices to monitor stress in real time. For example, [6] developed a stress recognition system based on heart sounds captured through in-ear devices, achieving results similar to traditional ECG recordings [6]. Similarly, in [7] the authors created a smartwatch equipped with Heart Rate Variability (HRV) sensors and sweat cortisol detection to track stress continuously and non-invasively. Another interesting approach comes from [8], where the authors combined vocal biomarkers with deep learning models to detect stress levels in employees. Along the same lines, authors in [9, 10] used LSTM neural networks to identify psychological disorders from voice signals, achieving a 98% accuracy rate in classifying stress. The integration of these technological advances with personalized mobile solutions opens the door to more accessible, continuous, and objective telemedicine tools.

In this context, this paper introduces a mobile application for the early detection and monitoring of work-related stress through voice analysis and machine learning. Unlike continuous monitoring systems, the app allows leaders to schedule periodic stress tests for collaborators, who provide short voice samples that are analyzed automatically. The model used in the application was selected after benchmarking more than 60 machine learning algorithms with the LazyPredict framework, trained on a dataset collected from real users.

The contributions of this work are: (i) the creation of a novel dataset specifically designed for work stress detection through voice; (ii) a systematic methodology for model selection and evaluation using the LazyPredict framework to systematically assess a wide range of algorithms; (iii) the design and implementation of a scalable microservices-based architecture that integrates mobile development technologies, low-latency APIs, and cloud-based machine learning services to support real-time stress inference; and (iv) the development of a mobile application that enables continuous monitoring and detection of work-related stress using voice input, providing a practical and accessible tool for end users.

The remainder of this paper is organized as follows: Sect. 2 reviews related work, Sect. 3 details the methodology and system design, Sect. 4 presents the results, and Finally, Sect. 5 concludes with future works.

2 Related Works

This section reviews key literature on stress detection using physiological signals, acoustic biomarkers, and machine learning. A Directed Literature Review (DLR) was conducted to ensure relevance, academic rigor, and minimal selection bias. The search, conducted in Scopus with the search string (“stress” AND (“voice” OR “biomarker” OR “machine learning” OR “detection”) AND “employees”), yielded 40 articles. Inclusion

criteria required studies to: (1) use physiological or acoustic signals for stress detection, (2) apply machine learning, and (3) include experimental validation. Exclusion criteria ruled out studies lacking machine learning or technical implementation. After screening, 10 articles were selected for in-depth review and categorized by primary input data type used for stress detection.

2.1 Models Based on Physiological Biomarkers

In [6], researchers evaluated heart sounds from in-ear devices compared to ECG signals, demonstrating that despite motion artifacts, in-ear HRV features processed via machine learning offer a promising non-invasive method for stress detection. Their best-performing model using in-ear data reached 81.1% accuracy, slightly surpassing the ECG-based system (76.5–78.9%). After applying an error-balanced data augmentation strategy to mitigate artifact bias, the model achieved ~77% accuracy, confirming the robustness of the approach for real-world biosignal monitoring. In [11], researchers introduced "Stress-Track," an IoT-based system that integrates body sensors for temperature, sweat, and motion monitoring with a stacked ensemble learning model combining Random Forest and Gradient Boosting. The system achieved 99.5% accuracy, classifying stress into three levels (low, normal, and high) and transmitting data to a cloud-based platform in real time. This work highlights the potential of IoT-connected wearables and multi-sensor fusion as effective, non-invasive tools for preventive stress monitoring in smart healthcare applications. Stress detection based on vocal biomarkers was explored in [8] using the ECAPA-TDNN model with Korean workers, yielding 77.5% accuracy and supporting the use of voice-based monitoring in workplace mental health contexts.

Researchers in [7], developed a smartwatch combining sweat cortisol sensors and HRV, using a hybrid GBM-RF model that achieved 100% accuracy, emphasizing the value of multimodal wearables in real-time physiological stress detection. In [12], researchers used AI-powered entertainment robots with multimodal emotion recognition to provide emotional support to students through adaptive interventions, illustrating the potential of personalized mental health technologies. Finally, in [9], researchers used deep learning models, including LSTM networks, applied to voice data from Indian patients to detect psychological stress, achieving 98% accuracy and demonstrating the method's applicability in low-resource clinical settings.

In summary, physiological and acoustic signal-based models, especially when integrated with IoT and AI, demonstrate high accuracy (often above 98%) and real-time monitoring capabilities. Nonetheless, challenges related to scalability, portability, and accessibility remain barriers to widespread implementation. For instance, scalability issues may arise when multiple users send audio samples simultaneously, potentially saturating the backend infrastructure. Portability is also a limitation, as performance may vary depending on the hardware capabilities of different mobile devices. Finally, accessibility remains a concern, since current solutions are primarily designed in a single language and may not adequately address the needs of users with hearing impairments or other disabilities.

2.2 Models Based on Acoustic Signals and Machine Learning

Researchers in [13] validated a machine learning–based respiratory-responsive vocal biomarker (RRVB) initially trained on voice samples from 1,700 asthma patients and healthy controls. The same model was tested on new data from patients with COVID-19, COPD, and other respiratory conditions recorded through smartphones. Using Logistic Regression on a 6-s sustained vowel ("ahh"), the model achieved 73.2% sensitivity and 62.9% specificity in distinguishing COVID-19 patients from healthy individuals. Results confirmed that acoustic features of the voice can serve as non-invasive indicators of respiratory impairment across diseases, languages, and geographic settings. Similarly, in [14] the researchers applied machine learning techniques to multiple non-invasive physiological signals—including ECG, EMG, GSR from the hands and feet, and respiration—to detect stress in car drivers. Their proposed AI-based Driver Assistance System (AI-DAS) followed a three-phase process of preprocessing, feature extraction, and classification using data from the drivedb dataset. Among several tested models, the Random Forest classifier achieved the best performance, reaching 98.2% accuracy, 97% sensitivity, and 100% specificity. The study demonstrates the effectiveness of multi-sensor integration and machine learning for real-time stress detection in dynamic and high-risk environments such as driving.

Voice analysis for early Parkinson's detection was explored in [15], where researchers used a public dataset of 252 samples and different models. Among these models, KNN, SVM, AdaBoost, ANN, and KDE, KNN performed best with 98.52% accuracy, supporting the use of vocal biomarkers and ML for non-invasive, cost-effective diagnosis. A hybrid AI model combining CNN, RNN, MKL, and MLP was also proposed for Parkinson's detection in [16], the researchers used 81 voice recordings and extracting acoustic features such as MFCCs, jitter, and shimmer. The model reached 91.11% accuracy.

In summary, machine learning has shown high potential for early disease detection using non-invasive signals, particularly in conditions like Parkinson's and COVID-19. However, stress detection remains relatively underexplored, highlighting a promising direction for future research. Although some of the studies reviewed, such as those focused on Parkinson's disease or COVID-19 detection, do not address stress directly, they were included due to their methodological relevance. These works demonstrate the effectiveness of combining acoustic biomarkers with machine learning for health-related monitoring tasks. Their inclusion provides evidence that the same methodological principles, voice signal processing and predictive modeling, can be extrapolated to stress detection in remote work contexts.

3 Methodology

This section outlines the proposed methodology for early detection and monitoring of work-related stress using voice analysis. The system is built on a scalable microservices architecture, integrating mobile technologies, low-latency APIs, and cloud-based machine learning. It covers dataset selection, model choice, evaluation metrics, and the system's logical and physical architecture.

3.1 Data Collection

A total of 60 participants (30 men, 30 women) between the ages of 18 and 50 (mean age = 20) were recruited for this study. Participants were categorized into two distinct groups based on their diagnostic status: one group consisting of individuals previously diagnosed with work-related stress, and a second group of individuals with no such diagnosis.

The dataset was constructed through a structured process involving data acquisition, labeling, preprocessing, and feature extraction. Voice samples were acquired using the WhatsApp messaging application. Participants were instructed to record an audio sample of approximately 15 s, prompted by either reading a short text or answering a simple question. Each collected voice recording was subsequently labeled according to the participant's self-reported emotional state at the time of the recording. This procedure yielded a dataset of 65 Spanish-language voice recordings, categorized as either "stressed" (n = 29) or "not-stressed" (n = 36).

The raw audio files were initially stored in the ".opus" format, which necessitated their conversion to the.wav format to ensure compatibility with the selected audio processing libraries. Following this conversion, an acoustic feature extraction process was performed to derive key biomarkers from each signal. The following 15 features were extracted based on [17, 18]:

- The first 13 Mel-Frequency Cepstral Coefficients (MFCCs)
- Pitch, representing the fundamental frequency of the voice
- Root Mean Square (RMS) energy, as a measure of audio intensity

The selected features represent key acoustic correlates of stress: spectral content, pitch, and intensity. The final dataset includes 16 columns: 13 MFCCs, pitch, RMS, and a binary class label (stressed/not stressed). We applied standardization due to the differing value ranges.

3.2 Model Selection

To select the best model for detecting work-related stress from voice data, we used LazyPredict, which automates training and evaluation of multiple classifiers, reporting metrics like accuracy, F1-score, and execution time.

For robust model validation, particularly given the limited size of our dataset, we implemented a Leave-One-Out Cross-Validation (LOOCV) strategy. LOOCV is highly effective for small datasets as it maximizes the utilization of available data for training in each iteration, thereby providing a nearly unbiased estimate of model generalization performance. This technique obviates the need for a static train-test split by iteratively using each data point once for testing while the remaining data serve as the training set. The comprehensive results of this systematic model comparison are presented in Table 2. Based on the initial evaluation in Table 1, the top five models selected for further analysis were: (i) DecisionTreeClassifier, which builds predictive models using hierarchical decision rules; (ii) LinearSVC, a linear classifier that maximizes the margin between classes; (iii) LogisticRegression, which estimates binary outcome probabilities from predictors; (iv) Perceptron, a basic neural network for binary classification; and

(v) RidgeClassifier, which applies ridge regression followed by thresholding to classify. Among these five candidates, the DecisionTreeClassifier demonstrated the superior performance in the comparative analysis, establishing it as the most promising model for this application.

Table 1. Results of the Top 5 Models Obtained with LazyPredict

Model	Accuracy	F1 Score	Time Taken
DecisionTreeClassifier	0.66	0.66	0.03
LinearSVC	0.66	0.66	0.02
LogisticRegression	0.65	0.65	0.04
Perceptron	0.63	0.63	0.03
RidgeClassifier	0.62	0.62	0.02

3.3 Evaluation Metrics

The model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. An exhaustive hyperparameter search was conducted for each of the five candidate models using GridSearchCV, with Leave-One-Out (LOO) cross-validation as the evaluation strategy. LOO, where each sample is used once as a validation fold, is well-suited for small datasets and provides robust performance estimates. The specific hyperparameters and their search spaces are shown in Table 2.

Table 2. Hyperparameters of the Top 5 Models

Model	Parameters
DecisionTreeClassifier	criterion, max_depth, min_samples_split
LinearSVC	C, loss, max_iter
LogisticRegression	C, penalty, solver
Perceptron	penalty, alpha
RidgeClassifier	Alpha

This tuning process allowed for the identification of optimal combinations in terms of accuracy and efficiency.

Table 3. GridSearchCV Results

Model	Accuracy	Precision	Recall
DecisionTreeClassifier	0.75	0.75	0.75

(*continued*)

Table 3. (*continued*)

Model	Accuracy	Precision	Recall
LogisticRegression	0.67	0.67	0.67
LinearSVC	0.66	0.66	0.66
Perceptron	0.65	0.65	0.65
RidgeClassifier	0.62	0.62	0.62

The Table 3 shows the results of the hyperparameter optimization process, confirmed the superiority of the DecisionTreeClassifier. This model achieved the best overall performance, demonstrating a balanced accuracy, precision, and recall of 0.75. This peak performance was obtained with the following hyperparameter configuration: a criterion of 'gini', a max_depth of 2, and a min_samples_split of 2.

The DecisionTreeClassifier markedly outperformed the other candidate models. In descending order of performance, the subsequent classifiers were LogisticRegression (0.67), LinearSVC (0.66), and Perceptron (0.65). The RidgeClassifier yielded the lowest performance among the top five, with a score of 0.62.

This outcome underscores the efficacy of hyperparameter optimization and solidifies the choice of the DecisionTreeClassifier as the definitive model for this application.

3.4 Logical Architecture of the Application

Fig. 1 presents the logical architecture of the mobile application, which outlines the system's components organized into four primary layers.

The Client Layer encompasses the mobile interface and defines two user roles: (i) Collaborators, employees who perform stress tests scheduled by a leader, and (ii) Leaders, who manage teams, invite collaborators, schedule weekly tests, and access historical results. If a collaborator records a "stressed" result three consecutive times, the system automatically alerts the leader. Inputs at this layer are user actions such as logging in and recording voice samples; outputs are data payloads (audio file, user ID, and authentication token) transmitted to the backend.

The Service Layer acts as the API gateway, receiving requests from the client, validating tokens, handling errors and logs, and passing authenticated data to the Business Logic Layer. For validation, it communicates with the User Database to confirm active sessions and retrieve user details.

The Business Logic Layer executes the core processing pipeline: (i) audio feature extraction (MFCCs, pitch, RMS), (ii) inference with the deployed Decision Tree model, (iii) binary stress classification (stressed / not stressed), and (iv) generation of tailored recommendations. Outputs include the classification result and feedback messages, while interactions with the Data Layer ensure that audio files are stored in the Voice Database and results in the Results Database.

Finally, the Data Layer (Persistence Layer) provides structured storage through three repositories: the User Database (accounts, roles, leader–collaborator mappings), the

Voice Database (raw audio samples with metadata), and the Results Database (classification outcomes linked to users and timestamps). This layer guarantees secure and persistent access to historical records for both collaborators and leaders.

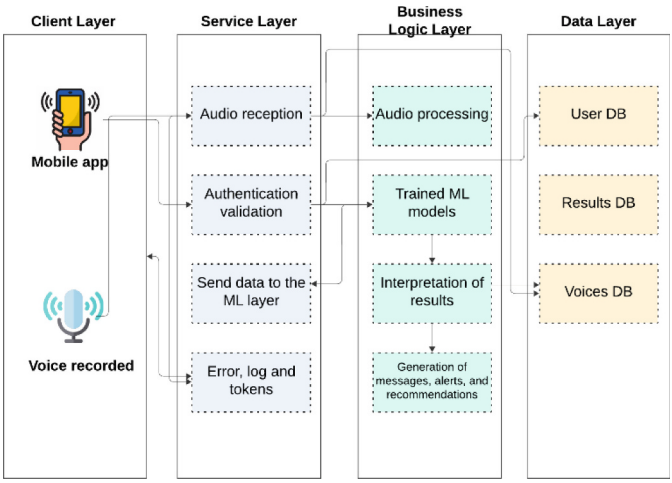


Fig. 1. Logical Architecture of the Application

3.5 Physical Architecture of the Application

The Fig. 2 shows the physical architecture of the application, designed to support a responsive and scalable real-time stress detection system. The user interaction is managed through a mobile application developed using the Flutter framework, which provides the interface for collaborators to record short audio clips.

Once captured, these recordings are transmitted via the internet to the backend infrastructure. This backend is built upon FastAPI [24], a modern web framework selected specifically for its high-performance asynchronous capabilities, seamless integration with machine learning models, and automated API documentation support. For data persistence, the system utilizes a PostgreSQL database, chosen for its robustness, native support for complex data types, and demonstrated reliability in digital health applications [25].

The core machine learning inference is handled by a dedicated service deployed on Amazon SageMaker. The workflow is orchestrated by the FastAPI backend that upon receiving an audio file, it forwards it to the Amazon SageMaker service. The deployed model processes the data, returns the predicted stress level, and this result is then relayed back through the backend to the mobile application for presentation to the user.

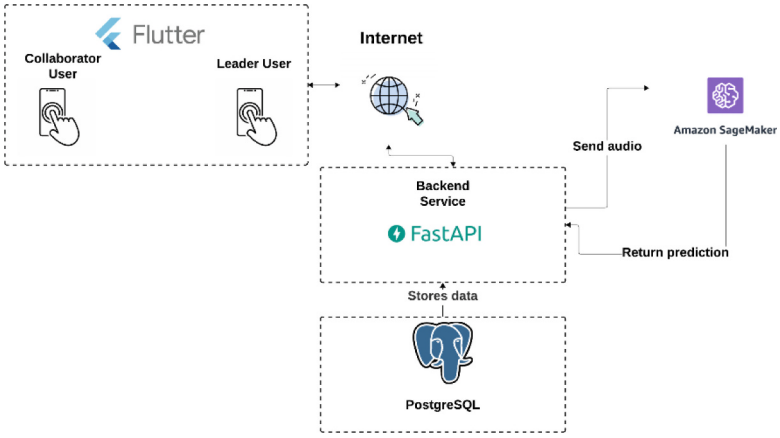


Fig. 2. Physical Architecture of the Application

4 Results

A key outcome of this work is a functional mobile application featuring an intuitive graphical user interface (GUI). The following figures present the main screens of the application prototype, illustrating core functionalities such as user login, audio recording, and the visualization of results and recommendations.

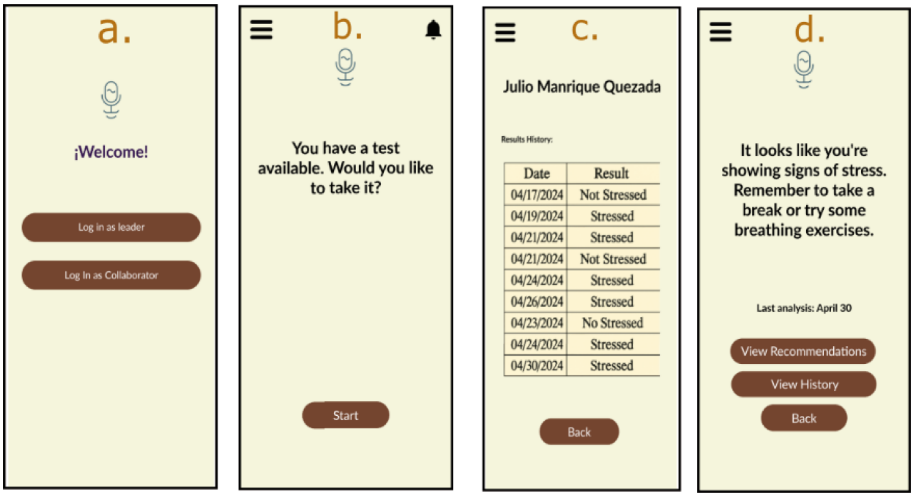


Fig. 3. Collaborator Voice Test Interface

Figure 3 shows the application user interface. The Fig. 3(a) shows the Panel of welcome screen, where users select their role: Leader (responsible for managing collaborators and scheduling stress tests) or Collaborator (end-user who performs voice-based

assessments). Figure 3(b) illustrates the stress test initiation screen, where collaborators can launch a pending test by pressing Start to begin the voice recording session. Figure 3(c) presents the results history, displayed in a tabular format with date and binary outcome (Stressed or Not Stressed), allowing both leaders and collaborators to monitor stress levels over time. Finally, Figure 3(d) depicts the results screen, which provides the outcome of the most recent analysis. When stress is detected, the interface displays an empathetic message encouraging self-care, along with navigation options: Recommendations, View History, and Back.

The application features a role-based design with two profiles: Leader and Collaborator. Upon login, users access their respective interfaces tailored to their roles. Leaders focus on team management and well-being monitoring, with access to historical stress data for collaborators and the ability to schedule weekly evaluations to foster preventive mental health practices. Collaborators conduct self-assessments by recording a short voice sample, which the system analyzes using acoustic biomarkers (MFCC, pitch, RMS) and a machine learning model to provide near real-time stress predictions. When stress is detected, collaborators receive actionable recommendations such as breathing exercises; if not, they can continue their tasks without interruption.

5 Conclusions and Future Work

This study presents the design, implementation, and validation of an end-to-end system for detecting work-related stress via machine learning analysis of acoustic biomarkers. Using the LazyPredict framework and Leave-One-Out cross-validation, a DecisionTreeClassifier was selected for its balance of accuracy, interpretability, and efficiency, enabling deployment on resource-limited mobile devices.

Our findings confirm that vocal features, MFCCs, pitch, and RMS energy, are reliable stress indicators, achieving 75% balanced accuracy, precision, and recall. This demonstrates the feasibility of objective, non-invasive periodic mental health monitoring scheduled by leaders.

The primary contribution is the integration of an effective predictive model with a functional mobile prototype, bridging theoretical research and practical application. This tool has significant potential in occupational health and telemedicine. An important limitation of this study is that the system does not perform continuous monitoring, but rather periodic assessments scheduled by leaders. This discrete approach may restrict its applicability for certain categories of remote workers, for example, professionals such as software developers who may not use their voice regularly while working. Moreover, since the tests are scheduled in advance, there is a potential risk of result manipulation, as individuals could alter their vocal patterns before a test (e.g., through medication or alcohol). These factors highlight the need for future work to explore more naturalistic and unobtrusive methods of stress detection, potentially through multimodal monitoring that integrates both acoustic and physiological signals. Future works includes expanding the dataset for improved generalizability, exploring deep learning architectures like CNNs or Transformers for enhanced feature extraction, and developing a multimodal system by combining vocal biomarkers with physiological signals from wearables to improve detection accuracy. Future evaluations should also incorporate user studies to

assess perceived usefulness, acceptance, and ethical considerations of voice-based stress monitoring in remote work environments.

Acknowledgments. "A la Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas por el apoyo brindado para realización de este trabajo de investigación a través del incentivo UPC-EXPOST-2025-2"

References

1. Adisa, T.A., Ogbonnaya, C., Adekoya, O.D.: Remote working and employee engagement: a qualitative study of British workers during the pandemic. Inf. Technol. People. **36**(5), 1835–1850 (2023). https://doi.org/10.1108/ITP-12-2020-0850
2. Perry, S.J., Carlson, D.S., Kacmar, K.M., Wan, M.M., Thompson, M.J.: Interruptions in remote work: a resource-based model of work and family stress. J. Bus. Psychol. **38**(5), 1023–1041 (2023). https://doi.org/10.1007/s10869-022-09842-y
3. Prasad, K.D.V., Vaidya, R., Rani, R.: Remote working and occupational stress: effects on IT-enabled industry employees in Hyderabad metro India. Front. Psychol. **14**, 1069402 (2023). https://doi.org/10.3389/fpsyg.2023.1069402
4. Magrizos, S., Roumpi, D., Georgiadou, A., Kostopoulos, I., Vrontis, D.: The dark side of meaningful work-from-home: a nonlinear approach. Eur. Manag. Rev. **20**(2), 228–245 (2023). https://doi.org/10.1111/emre.12534
5. Kalmanovich-Cohen, H., Stanton, S.J.: When working from home leads to burnout: the role of work–family conflict, job stress and partner work practices. Int. J. Manpow. **46**, 1321 (2025). https://doi.org/10.1108/IJM-11-2024-0773
6. Benesch, D., Villatte, B., Vinet, A., Hébert, S., Voix, J., Bouserhal, R.E.: Stress classification with in-ear heartbeat sounds. Comput. Biol. Med. **186**, 109555 (2025). https://doi.org/10.1016/j.compbiomed.2024.109555
7. Ding, C., Zhang, Y., Ding, T.: A systematic hybrid machine learning approach for stress prediction. Peer J Comput. Sci. **9**, e1154 (2023). https://doi.org/10.7717/PEERJ-CS.1154
8. Namkung, J., et al.: Novel deep learning-based vocal biomarkers for stress detection in Koreans. Psychiatry Investig. **21**(11), 1228–1237 (2024). https://doi.org/10.30773/pi.2024.0131
9. Kumar, A., Shaun, M.A., Chaurasia, B.K.: Identification of psychological stress from speech signal using deep learning algorithm. E-Prime – Adv. Electr. Eng. Electron. Energ.s. **9**, 100707 (2024). https://doi.org/10.1016/j.prime.2024.100707
10. Bhangale, K., Kothandaraman, M.: Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. Electronics (Switzerland). **12**(4), 839 (2023). https://doi.org/10.3390/ELECTRONICS12040839
11. Al-Atawi, A.A., et al.: Stress monitoring using machine learning, IoT and wearable sensors. Sensors. (Basel, Switzerland). **23**(21), 8875 (2023). https://doi.org/10.3390/s23218875
12. Lu, Z., Zhou, L.: Entertainment robots based on internet of things and artificial intelligence in the process of student stress and emotion recognition. Entertainment Comput. **52**, 100786 (2025). https://doi.org/10.1016/j.entcom.2024.100786
13. Kaur, S., et al.: Development and validation of a respiratory-responsive vocal biomarker–based tool for generalizable detection of respiratory impairment: independent case-control studies in multiple respiratory conditions including asthma, chronic obstructive pulmonary disease, and COVID-19. J. Med. Internet Res. **25**, e44410 (2023). https://doi.org/10.2196/44410

14. Siam, A.I., Gamel, S.A., Talaat, F.M.: Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. Neural Comput. & Applic. **35**(17), 12891–12904 (2023). https://doi.org/10.1007/s00521-023-08428-w
15. Kumar, A., Singh, J.P., Paygude, P., Daimary, R., Prasad, S.: Advanced comparative analysis of machine learning algorithms for early Parkinson's disease detection using vocal biomarkers. Digital Health. **11**, 20552076251342878 (2025). https://doi.org/10.1177/20552076251342878
16. Shen, M., Mortezaagha, P., Rahgozar, A.: Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. Sci. Rep. **15**(1), 11687 (2025). https://doi.org/10.1038/S41598-025-96575-6
17. Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., Poria, S.: A review of deep learning techniques for speech processing. Inf. Fusion. **99**, 101869 (2023). https://doi.org/10.1016/j.inffus.2023.101869
18. Dahl, K.L., Stepp, C.E.: Effects of cognitive stress on voice acoustics in individuals with Hyperfunctional voice disorders. Am. J. Speech Lang. Pathol. **32**(1), 264–274 (2023). https://doi.org/10.1044/2022_AJSLP-22-00204
19. Lemenkova, P.: Improving bimonthly landscape monitoring in Morocco, North Africa, by integrating machine learning with GRASS GIS. Geomatics. **5**(1), 5 (2025). https://doi.org/10.3390/GEOMATICS5010005
20. Alnajim, R.A., Fakieh, B.: A tourist-based framework for developing digital Marketing for Small and Medium-Sized Enterprises in the tourism sector in Saudi Arabia. Data. **8**(12), 179 (2023). https://doi.org/10.3390/DATA8120179
21. Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A., Blokland, M.: The effects of data balancing approaches: a case study. Appl. Soft Comput. **132**, 109853 (2023). https://doi.org/10.1016/J.ASOC.2022.109853
22. Bhatt, C.M., Patel, P., Ghetia, T., Mazzeo, P.L.: Effective heart disease prediction using machine learning techniques. Algorithms. **16**(2), 88 (2023). https://doi.org/10.3390/A16020088
23. Elshewey, A.M., Shams, M.Y., El-Rashidy, N., Elhady, A.M., Shohieb, S.M., Tarek, Z.: Bayesian optimization with support vector machine model for Parkinson disease classification. Sensors. **23**(4), 2085 (2023). https://doi.org/10.3390/S23042085
24. Odofin, O.T., Adekunle, B.I., Ogbuefi, E., Ogeawuchi, J.C., Adanigbo, O.S., Gbenle, T.P.: Improving healthcare data intelligence through custom NLP pipelines and fast API micro services. J. Front. Multi. Res. **4**(1), 390–397 (2023). https://doi.org/10.54660/.JFMR.2023.4.1.390-397
25. Amardeep, R., Soumya, H.S., Srisha Raghavendra, H.S., Gowda, S.S., Jain, Y.R.: Advancement in health monitoring through IOT and post-gre SQL analytics. In: International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications, ICETCS 2024 (2024). https://doi.org/10.1109/ICETCS61022.2024.10543622

Evaluation of Machine Learning Algorithms to Estimate the Prevalence of Anemia in Children at the District Level in Peru

Anthony Mamani[1], Daniel A. Gutierrez-Pachas[1(✉)], and Luis Rosero[2]

[1] Departamento de Computación, Universidad Católica San Pablo, Arequipa, Peru
{anthony.mamani,dgutierrezp}@ucsp.edu.pe

[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil
lrosero@usp.br

Abstract. An adequate assessment of public health indicators and metrics is crucial for national governments, enabling them to formulate more efficient public policies that benefit society. Traditionally, statistical methods were employed to obtain spatial estimates in small areas with limited or no information, a technique known as Small Area Estimation (SAE). Despite the success of these statistical techniques, they present challenges when dealing with increasing dimensionality and volume of information. For this reason, evaluating modern techniques is crucial to assessing SAE problems. We address an alternative to traditional methods for SAE problems using Machine Learning (ML) algorithms. To validate our results, we utilize the estimation of childhood anemia prevalence at the district level in Peru, based on real data, and compare it with the statistical method, spatial Fay-Herriot. Our evaluation uses regression and autocorrelation spatial metrics, as well as consistency with real data. Our results demonstrate that ML algorithms can offer competitive, and in some cases, superior, predictive capabilities despite data scarcity and inconsistencies. Finally, we conclude that ML algorithms are a viable alternative for governments to establish more efficient public policies.

Keywords: Small Area Estimation · Machine Learning · Childhood Anemia · Public Health

1 Introduction

Accurate estimation of public health indicators at specific geographic scales is vital for designing effective and fair policy interventions. In situations where direct survey estimates are either unreliable or unavailable, statistical analysis techniques are employed with relative success. Childhood anemia is a critical public health problem in developing countries, and obtaining reliable estimates of anemia rates in small areas is a significant challenge. If not adequately addressed,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 480–494, 2026.
https://doi.org/10.1007/978-3-032-20322-9_33

anemia can progress to a chronic condition with serious consequences for growth, cardiac function, and cognitive development. These effects may also be influenced by underlying conditions contributing to anemia [1]. According to the *National Institute of Statistics and Informatics* (INEI), the data of [12] show that in 2019, 40.1% of children aged 6 to 35 months and 29.5% of children under 5 years of age suffered from anemia. The highest percentages were recorded in higher-altitude regions, such as Puno (69.9%). Similar cases occur in Cusco and Huancavelica, which exceed 54% in both cases.

Peruvian public health institutions require current, more accurate, and spatially disaggregated data and employ the *National Demographic and Family Health Survey* (ENDES) to obtain this information. However, ENDES has limitations in providing data at the necessary local scale. In Peru, the district (in Spanish *distrito*) is the third-level administrative division, subordinate to provinces and departments. 1,874 districts nationwide represent the primary unit for local governance and the implementation of public health actions. It is precisely at this satisfactory geographic resolution that ENDES, due to its sampling design, is often unable to produce reliable estimates. This lack of robust district-level data hinders the development of localized, evidence-based policies, potentially leaving the most vulnerable populations behind. Anemia estimation in Peru has recently been addressed in [19], as an SAE problem using the spatial *Fay-Herriot* (SFH) method, and a comprehensible variation was made to the traditional model by adding height to the conventional geospatial variables (longitude and latitude). Although these results were novel in the Peruvian context, they only considered 14 predictor variables out of the total real data obtained from ENDES in 2019. Meanwhile, ML algorithms that allow their efficiency to be evaluated by incorporating more variables into their analysis have not been tested.

We evaluate five ML algorithms for generating robust and reliable district-level estimates of childhood anemia prevalence across Peru. The performance of several algorithms—including *Random Forest* (**RF**), *XGBoost* (**XGB**), *Support Vector Regression* (**SVR**), *Multilayer Perceptron* (**MLP**), and a *Stacking Ensemble* (**SE**)—is rigorously evaluated. Additionally, we trained the algorithms on a set of 111 census variables and three geographic variables, directly comparing their performance based on regression metrics, bootstrap-derived uncertainty estimates, and spatial correlation with the classic SFH method. However, although the algorithms perform acceptably well when evaluating computational metrics, this is not decisive because there is little confidence in the quality of the census data, which differs from estimations by the *Peruvian Ministry of Health* (MINSA). For this reason, our proposal includes an external validation analysis based on another socioeconomic metric, such as the Human Development Index (HDI), a composite indicator that summarizes key dimensions of human development—including health, education, and income—at the district level. Through this approach, our research seeks to validate the effectiveness of these modern, data-driven techniques and establish a robust methodological framework for government institutions to develop more targeted and equitable

public health interventions. Finally, we orient our proposal towards the following research questions:

- Which ML algorithm has the best performance in the observed districts?.
- Which ML algorithm has the best performance in the unobserved districts?.
- Which ML algorithm is most consistent with real data?.

2 Related Works

Making accurate and reliable estimates in small areas is a significant challenge with broad applicability, [14]. However, statistical methods are often unable to address problems in the context of high dimensionality and large data volumes. This approach leverages "*borrowed strength*" from external data, enabling the production of more representative and applicable results. Although statistical methods have traditionally dominated SAE problems, research also addresses mixed approaches, including linear, Bayesian, spatial, and hierarchical models, as noted by [20].

More recently, ML algorithms have emerged as a novel and powerful alternative for SAE. The application of tree-based models, in particular, has gained significant traction. Pioneering work in the related field of poverty mapping has demonstrated the practical advantages of these methods, utilizing classification trees [2] and later regression trees [3] to estimate complex poverty measures, such as incidence, gap, and severity. Building on this, a more formal framework has been proposed for integrating tree-based models with mixed effects. Krennmair and Schmid [16] introduced a coherent framework based on mixed-effects random forests for estimating small-area averages, proposing a non-parametric bootstrap estimator to assess the uncertainty of these estimates—a technique central to our own methodology. This line of research continues to evolve rapidly. Recent extensions have adapted these models for specific data types, such as the generalized mixed effects random forest for estimating poverty rates from binary response variables in Mexico [6]. Further adaptations for count outcomes also explore various bootstrap methodologies for uncertainty assessment [7]. Research in other contexts, such as in African countries, has also successfully applied ML techniques to estimate health indicators, with some studies concluding that RF presents superior computational performance [15,18,21,24].

Despite the growing interest and rapid development of specialized ML models for SAE, there remains a notable lack of comprehensive research directly comparing the performance of diverse statistical models with various ML algorithms in SAE. However, some pioneering studies demonstrate the potential of ML; for example, [22] compared XGB with a structured additive regression model to estimate public health indicators in the Netherlands, concluding that ML algorithms can be a robust alternative offering advantages in predictive accuracy and ease of implementation. In the specific Peruvian context, existing approaches for estimating childhood anemia are primarily based on statistical methods. For instance, [19] combined ENDES 2019 data with National Census data, using

14 predictor variables within a Fay-Herriot model framework to estimate childhood anemia prevalence in Peru. This highlights a critical gap that our study aims to address by rigorously evaluating a suite of modern ML algorithms for district-level anemia estimation.

3 Methodology

In this section, we begin by explaining the data preprocessing. Next, we discuss the ML algorithms, and finally, we examine the computational metrics used.

3.1 Data Preprocessing

We used real data from national sources, with the 2019 ENDES survey providing the target variable: district-level prevalence of childhood anemia [12]. In that year, only 836 of Peru's 1,874 districts (44.61%) included at least one valid anemia observation. This notable data sparsity highlights the limitations of relying solely on direct survey estimates and underscores the need for model-based methods. In districts with available data, the outcome variable was defined as the proportion of anemia-positive cases among the sampled children, yielding a continuous value between 0 and 1 suitable for model training.

As predictors, we relied primarily on the 2017 National Census, which provides comprehensive information on population, housing, and indigenous communities [11]. These data were accessed through the REDATAM system [10], which allows customized queries by thematic category (e.g., Housing, Population, Households) and geographic level (department, province, and district). Because many census questions included a large number of categories, directly using them would have led to severe high-dimensionality issues. To address this, we applied a systematic feature engineering process consisting of two key steps: *conceptual grouping* and *normalization*.

For example, the census question "*Type of housing*" contained over ten distinct response categories. We grouped them into four coherent synthetic variables: "*Formal Housing*" (detached houses, apartments, or multi-unit buildings), "*Informal Housing*" (slum dwellings, huts, improvised structures), "*Adapted Non-Residential Dwellings*", and "*Other Housing Types*". For each district, we computed the absolute number of dwellings in each new category and then normalized these counts by the total number of homes in the district. This two-step process was systematically applied to all selected census questions, producing a standardized matrix of socioeconomic predictors. The final census-based feature set included 111 variables.

In addition, we incorporated three geographic predictors: altitude, latitude, and longitude. Altitude values were obtained from [4], while latitude and longitude were extracted from the official shapefile of Peruvian districts provided by GeoGPSPerú [8]. Geographic variables present challenges due to their differing scales (degrees vs. meters) and potential correlations, which may affect computational performance [17]. To ensure consistency across models, we standardized these variables using a *StandardScaler*.

In total, our feature matrix consisted of 114 predictors: 111 census-derived socioeconomic variables and 3 geographic variables. This high-dimensional feature space, in contrast to benchmark studies that employed only 14 variables [19], was designed to allow modern ML algorithms to capture more complex and nuanced patterns in the determinants of childhood anemia.

A crucial aspect of our proposal is the data partitioning strategy for model training, which deviates from conventional partitioning schemes (70/30 training-testing split). As this survey provides direct estimates for only 44.61% of districts, a standard random split is infeasible. Therefore, the algorithms are trained exclusively on the observed data set—that is, the subset of districts for which ENDES provides at least one sample. Once trained, the models are used to generate predictions for the entire country, covering both observed and unobserved districts. The primary challenge, therefore, lies in evaluating the performance of these nationwide estimates. While the direct projections from the observed districts constitute the training data, their high intrinsic variance makes them an imperfect benchmark for simple validation.

3.2 ML Algorithms

We selected a diverse set of supervised machine learning models. The chosen algorithms include: two tree-based methods (RF and XGB), a kernel-based approach (SVR), a neural network (MLP), and a final SE model designed to leverage the strengths of the individual models. We chose RF for its robustness to outliers, its ability to handle high-dimensional data without extensive feature selection, and its inherent capacity to capture complex interaction effects between variables. Also, XGB is another robust tree-based ensemble algorithm known for its state-of-the-art performance on tabular data [5]. Unlike RF, which builds trees independently, XGB builds them sequentially, with each new tree correcting the errors of the previous one. It was selected for its high predictive accuracy, computational efficiency, and built-in regularization techniques that help prevent overfitting. SVR was included because it represents a different family of algorithms (kernel methods), capable of modeling complex, non-linear relationships through the use of a kernel function (in our case, a Radial Basis Function or RBF). MLPs are universal approximators, capable of learning highly complex, non-linear functions [9] that can uncover deeper patterns in the data that other models may miss. and utilize SE to improve the predictive performance [23].

All models were implemented using the `scikit-learn` ecosystem in Python. To ensure reproducibility and to prevent information leakage, each algorithm was wrapped in a `Pipeline` that includes the StandardScaler function for feature standardization. Hyperparameters were selected by grid search with 5-fold cross-validation (`GridSearchCV`); the best configurations are summarized in Table 1. For models sensitive to the output range, particularly the MLP, we employed a `TransformedTargetRegressor` function to wrap the model. This applied a logit transform to the prevalence data before training and an inverse-logit (sigmoid) transform to the predictions, ensuring they remained within the valid range and

stabilizing the learning process. After hyperparameter selection, the final optimized estimator for each method was re-fitted on the entire set of 836 observed districts before being used for nationwide prediction.

Table 1. Optimal hyperparameters for each ML algorithm.

ML algorithm	Optimal hyperparameters
RF	`n_estimators`=200, `max_depth`=12, `max_features`=0.5, `min_samples_leaf`=2, `min_samples_split`=8
XGB	`n_estimators`=400, `learning_rate`=0.01, `max_depth`=3, `subsample`=0.6, `colsample_bytree`=0.7, `gamma`=0, `reg_alpha`=0, `reg_lambda`=1
SVR	`kernel`='rbf', `C`=5, `epsilon`=0.15, `gamma`=0.01, `shrinking`=True, `tol`=0.0001
MLP	`hidden_layer_sizes`=(128, 64, 32), `activation`='relu', `alpha`=0.001, `learning_rate`='constant', `learning_rate_init`=0.005, `solver`='adam', `early_stopping`=True, `validation_fraction`=0.1, `tol`=0.0001, `max_iter`=200

3.3 Evaluation Metrics

A critical component of any SAE framework is the quantification of uncertainty, as point predictions alone are insufficient for supporting informed policymaking. Since analytical derivation of uncertainty is often intractable for complex ML models, we employed a non-parametric bootstrap procedure to estimate the Mean Squared Error (MSE) and construct confidence intervals. This resampling method is widely accepted and aligns with recent SAE literature that combines ML predictors with bootstrap-based uncertainty estimation [16]. The procedure involved generating 500 bootstrap samples by resampling with replacement from the original 836 districts with observed data. For each sample, the final, fully tuned model was retrained from scratch and used to predict anemia prevalence for all 1,874 districts. This yielded an empirical distribution of 500 predictions per district, forming the basis for our uncertainty metrics.

To rigorously evaluate and compare the performance of the machine learning models, we developed an evaluation framework that addresses both the districts with direct survey estimates and those without. The performance of our ML models was also benchmarked against the SFH model estimates from a recent

related study [19]. For the 836 districts with available ENDES data, we assessed predictive accuracy using standard metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (RÂš). We also computed the Mean Bias to assess systematic over- or under-prediction. Beyond point-estimate accuracy, we evaluated the quality of the uncertainty intervals generated by our bootstrap procedure. The 95% coverage probability was calculated as the proportion of districts where the direct survey estimate fell within the model's 95% bootstrap confidence interval (derived from the 2.5th and 97.5th percentiles); a well-calibrated model should yield coverage close to 0.95. Complementing this, we computed the average confidence interval length as a measure of the precision of the estimates.

For the 1,038 districts without direct survey estimates, where traditional predictive accuracy metrics cannot be computed, our evaluation focused on the characteristics and plausibility of the predictions. We analyzed descriptive statistics of the out-of-sample estimates, including the overall mean prediction, the prediction range (minimum and maximum), the average standard deviation derived from the bootstrap distributions, and the average length of the 95% confidence intervals. These metrics provide crucial insights into the stability, range, and overall uncertainty of the model-based estimates in domains where no direct information is available. To further validate our results, we conducted two additional analyses. First, we computed Moran's I statistic (using both Queen and K-Nearest Neighbors weight matrices) on the model residuals to assess for any remaining spatial autocorrelation. Second, we performed an external validation by calculating the Spearman rank correlation between our district-level anemia estimates and the IDH, where a negative correlation is expected, providing a measure of the external validity of our model's predictions.

4 Results

4.1 Model Performance in Observed Districts

The performance of the five ML models was evaluated against the direct survey estimates for the 836 districts with available data. The results were also benchmarked against the point estimates from the SFH model proposed by [19]. A comprehensive summary of the performance metrics is presented in Table 2.

The evaluation of predictive accuracy reveals a clear hierarchy among the models. The XGB model achieved the highest performance in terms of predictive accuracy, with the lowest RMSE (0.0541) and the highest R^2 (0.9111). However, the assessment of uncertainty quantification highlights a critical trade-off. The MLP was the only model to achieve near-nominal 95% coverage (0.994), but this came at the cost of precision, yielding the widest average confidence intervals (0.2837). At the other extreme, the RF and SVR models performed poorly in uncertainty estimation, with coverage rates below 33%, rendering their narrow confidence intervals unreliable.

Based on this in-sample analysis, the SE model is identified as the most promising candidate due to its superior balance across all evaluation criteria.

Table 2. Performance comparison of ML and SAE model (SFH) in the 836 districts with survey data.

Model	RMSE	MAE	Bias	R^2	95% Coverage	Avg. 95% CI Length
RF	0.0867	0.0678	0.0006	0.7717	0.3230	0.1495
XGB	**0.0541**	**0.0424**	0.0006	**0.9111**	0.8254	0.2471
SVR	0.1236	0.1053	0.0136	0.5368	0.3194	**0.1304**
MLP	0.0749	0.0592	−0.0020	0.8298	**0.9940**	0.2837
SE	0.0687	0.0535	**−0.0001**	0.8569	0.8589	0.1907
SFH	0.1030	0.0776	−0.0075	0.6782	–	–

While the XGB model achieved slightly higher predictive accuracy, its coverage rate (0.8254) is lower than that of the SE model (0.8589). Conversely, while the MLP offered near-perfect coverage, it did so with significantly wider confidence intervals (0.2837 vs. SE's 0.1907) and a lower R^2 (0.8298 vs. SE's 0.8569). Therefore, the SE model represents the most effective compromise, delivering high predictive accuracy (R^2 of 0.8569) while maintaining a more reliable and precise quantification of uncertainty than its closest competitors. This balance is critical for providing robust estimates, pending the analysis in unobserved districts.

4.2 Model Performance in Unobserved Districts

For the 1,038 districts where direct survey estimates are unavailable, the evaluation focused on the characteristics and plausibility of the model-based predictions. The statistical properties of these out-of-sample estimates for each model are detailed in Table 3.

Table 3. Statistical properties of model-based predictions for the 1,038 districts without direct survey estimates.

Model	Mean Prediction	Min Prediction	Max Prediction	Avg. SD	Avg. 95% CI Length
RF	0.4363	0.2334	0.6065	**0.0285**	**0.1105**
XGB	0.4348	0.2182	0.6301	0.0811	0.3128
SVR	0.4509	0.2448	0.6563	0.0372	0.1429
MLP	0.4303	0.0939	0.7484	0.0954	0.3659
SE	0.4388	0.2103	0.6318	0.0465	0.1808
SFH	0.4155	0.2184	0.7237	–	–

Note: Bold values indicate the lowest (best) uncertainty metrics. However, in this case, the narrow confidence intervals of RF and SVR reflect overconfident estimates with poor coverage (as shown in the observed districts), rather than genuinely superior performance.

A notable finding is the high consistency in the mean predicted prevalence across all ML models, which ranges narrowly from 0.4303 (MLP) to 0.4509 (SVR). This consensus suggests that the models captured a stable underlying signal from the predictor variables. These mean estimates are also consistently higher than those of the SFH benchmark model (0.4155).

The analysis of prediction ranges and uncertainty metrics for these unobserved districts reinforces the conclusions drawn from the in-sample evaluation. The MLP model exhibits the widest prediction range (from a minimum of 0.0939 to a maximum of 0.7484) and, consistent with its in-sample behavior, yields the highest average standard deviation (0.0954) and the most extended average confidence interval (0.3659). This refers to the fact that the MLP is capable of making more extreme predictions while correctly quantifying its higher uncertainty when extrapolating to new areas. Conversely, the RF and SVR models again display signs of overconfidence, producing the narrowest average confidence intervals (0.1105 and 0.1429, respectively). This finding aligns with their poor coverage performance in the observed districts.

These out-of-sample results further support the selection of the SE model as the most robust choice. Its average confidence interval length (0.181) occupies a sensible middle ground, avoiding the extreme overconfidence of the RF and SVR models and the high imprecision of the MLP. The SE model's ability to maintain a balanced and stable uncertainty profile in areas without ground-truth data confirms its reliability, making it the preferred model for generating the final district-level anemia prevalence map for Peru.

4.3 Consistency with Real Data

To provide visual and external validation for our model's estimates, we conducted a two-level spatial comparison. First, we compared our results against the official 2019 anemia prevalence statistics reported by the MINSA [13].

Since MINSA only provides data at the department level, our district-level predictions from the SE model were aggregated by calculating a simple average of the district estimates within each department, enabling a direct comparison. It is crucial to emphasize that the MINSA estimates were not used at any stage of model training or tuning, serving exclusively as an independent benchmark for external validation. As shown in Fig. 1, a visual comparison at the department level reveals a general concordance in identifying broad regional patterns. Both the official MINSA data and our aggregated SE estimates correctly identify higher prevalence rates in the southern Andean regions and lower rates along the coast. However, the comparison also highlights significant discrepancies. The most notable divergence is observed in the Puno department, where MINSA reports a prevalence substantially higher than our model estimates. Similar, though less pronounced, differences are also apparent in Cusco and some northern regions. Although the department-level view is helpful for external validation, the primary strength of our approach lies in its high spatial resolution.

Figure 2 illustrates this by providing a detailed district-level view of the Arequipa department. The map of direct ENDES estimates (a) illustrates the sparse

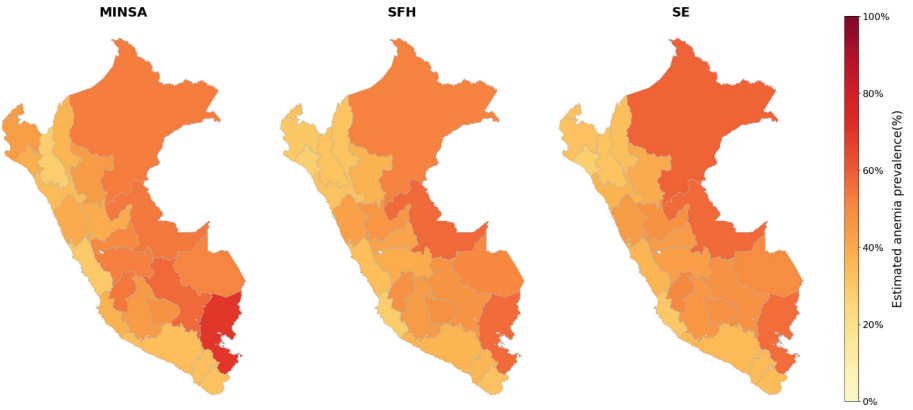


Fig. 1. Comparison of estimated anemia prevalence by Peruvian Departments: (a) Real Data (MINSA), (b) SFH method, and (c) SE algorithm.

nature of the survey data, with many districts lacking information. In contrast, both the SFH (b) and our SE model (c) provide a complete map. However, the SE model reveals significant intra-departmental heterogeneity that is completely masked in aggregated statistics, showcasing its ability to generate critical granular insights for targeted local interventions. These visual analyses confirm the model's ability to capture macro-level trends while offering crucial high-resolution detail.

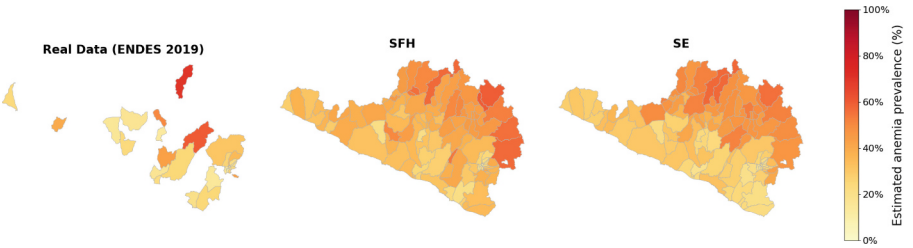


Fig. 2. Comparison of estimated anemia prevalence in the Arequipa department for (a) Real Data (ENDES 2019), (b) SFH model, and (c) SE algorithm.

The discrepancies with official statistics (MINSA) likely stem from methodological differences (e.g., sampling frames, weighting adjustments). Although the opaque nature of the official estimation methodology precludes a direct reconciliation, our explicit, transparent, and replicable ML-based approach provides a valuable, data-driven alternative for policy analysis and planning.

To quantify the degree of spatial clustering in the estimates, we calculated the statistic I_M for each model. To ensure robustness, we used two distinct spatial weights matrices: a Queen contiguity matrix, which defines neighbors

based on shared borders, and a K-Nearest Neighbors (KNN, k = 5) matrix, which overcomes the "*island*" problem (e.g., for districts like Amantaní) by assigning neighbors based on proximity and shows the results in Table 4.

Table 4. Moran's I_M statistic for anemia prevalence estimates using Queen and KNN spatial weights matrices.

Model	I_M (Queen)	I_M (KNN)	p-value
RF	0.716	0.735	< 0.001
XGB	0.618	0.633	< 0.001
SVR	0.603	0.611	< 0.001
MLP	0.580	0.594	< 0.001
SE	0.680	0.697	< 0.001
SFH	0.684	0.695	< 0.001

Table 4 confirms that all models generated estimates with strong and statistically significant ($p < 0.001$) positive spatial autocorrelation. A key finding is that our SE model achieves Moran's I values that are almost identical to those of the SFH benchmark. This is particularly noteworthy because, while the SFH model is explicitly designed to model geospatial correlations, our SE model achieves comparable performance by treating latitude, longitude, and altitude as additional predictor variables. Although the RF model yielded slightly higher Moran's I values, its previously identified deficiencies in uncertainty estimation (i.e., severe under-coverage) make it a less reliable choice. Therefore, the ability of the SE model to implicitly capture complex spatial patterns almost as effectively as a dedicated spatial model demonstrates its robustness and makes it highly competitive against reference methods in SAE.

To assess the real-world plausibility of the estimates, we conducted an external validation by correlating the model-generated anemia prevalence with the district-level HDI, a robust indicator of socioeconomic conditions. We deliberately excluded the final HDI score from the model training process. We acknowledge that many of the 114 census-based predictors (e.g., "Formal Housing") are components of, or are highly correlated with, the socioeconomic dimensions captured by the HDI. Therefore, this analysis is not presented as a validation against a completely independent variable, but rather as a critical test of socioeconomic consistency. A successful model should learn the relationship between these predictors and anemia in a way that aligns with the well-established inverse correlation between development and anemia. The strength and consistency of this correlation thus serve as a valuable measure of a model's real-world plausibility and generalizability.

In Fig. 3, each point represents a district, colored according to the official anemia prevalence thresholds set by MINSA: green (<20%, no problem), yellow (20–30%, mild), orange (30–40%, moderate), and red (>40%, severe), and reveals

that the SE model and the SFH model both display a strong and consistent negative trend, indicating their overall predictions are highly comparable.

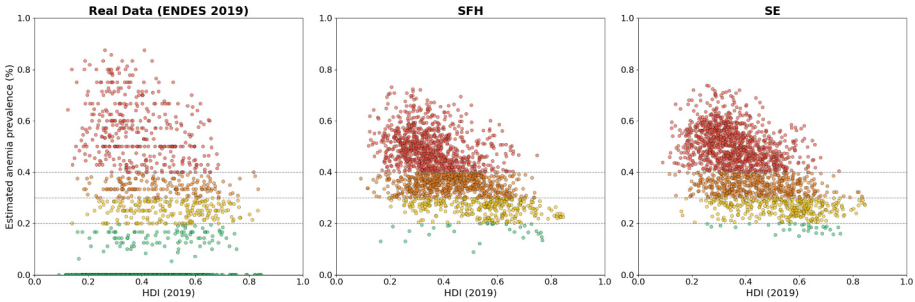


Fig. 3. Comparison of the relationship between HDI and estimated anemia prevalence for (a) Real Data (ENDES 2019), (b) SFH model, and (c) SE algorithm.

The most notable difference lies in the distribution of districts classified in the "no problem" category (represented by green points). The SFH model exhibits greater dispersion for this group, with low-anemia predictions scattered across a wide range of HDI values, including in areas with lower development, where such outcomes would be less expected. This suggests that the SE model generates more consistent predictions that align more closely with the socioeconomic gradient measured by the HDI.

Table 5. Pearson's correlation coefficient between district-level HDI and predicted anemia prevalence for each model.

Model	Pearson's Correlation - r	p-value
RF	-0.5443	< 0.0001
XGB	-0.4878	< 0.0001
SVR	-0.4766	< 0.0001
MLP	-0.4365	< 0.0001
SE	-0.5242	< 0.0001
SFH	-0.4891	< 0.0001

Nevertheless, the overall dispersion of estimates for both models is very similar, demonstrating that the performance of the SE is highly competitive and comparable to that of the dedicated spatial benchmark. The results in Table 5 confirm these visual findings. All models produce a statistically significant negative correlation, consistent with the epidemiological hypothesis. The RF model achieved the strongest correlation ($r = -0.5443$), followed very closely by our

selected SE model ($r = -0.5242$). Crucially, both of these top-performing ML models showed a stronger correlation with the HDI than the SFH benchmark ($r = -0.4891$). This suggests that by leveraging the richer, 114-variable feature space, the ML models successfully capture a more nuanced and accurate picture of the socioeconomic reality than the SFH model. This finding reinforces that the SE approach not only fits the survey data well and provides balanced uncertainty but also generates estimates that are significantly more consistent with the independent socioeconomic landscape of Peru, making it a more reliable tool for policy-making.

5 Conclusions

This study demonstrates that ML algorithms, when integrated into a robust statistical framework, represent a powerful alternative for SAE. We address a primary limitation of standard ML models—their lack of inherent uncertainty quantification—by employing a non-parametric bootstrap framework.

This hybrid ML-Bootstrap approach elevates ML algorithms from simple predictive tools to a comprehensive estimation framework, making them directly comparable to established methods, such as the SFH. Within this framework, our evaluation identified the SE model as the most effective, striking a superior balance between high predictive accuracy and reliable uncertainty quantification.

Furthermore, the SE model proved highly competitive against the SFH benchmark in terms of spatial coherence, without explicitly modeling geospatial dependencies. We conclude that the ability of ML to leverage high-dimensional data, when properly equipped to estimate uncertainty, generates more precise and geographically consistent estimates for critical public health indicators.

This work provides Peruvian health institutions with a validated, transparent, and replicable methodological framework to generate the granular evidence needed for more effective and targeted anti-anemia policies. Our findings firmly establish this hybrid ML approach as a powerful tool for evidence-based governance. Looking ahead, future research should aim to integrate more diverse data sources to enhance predictive power further and continue the development of robust uncertainty quantification methods tailored for complex ML models, thereby improving their utility for informing public policy.

References

1. Allali, S., Brousse, V., Sacri, A.S., Chalumeau, M., de Montalembert, M.: Anemia in children: prevalence, causes, diagnostic work-up, and long-term consequences. Expert Rev. Hematol. **10**(11), 1023–1028 (2017)
2. Bilton, P., Jones, G., Ganesh, S., Haslett, S.: Classification trees for poverty mapping. Comput. Stat. Data Anal. **115**, 53–66 (2017). https://doi.org/10.1016/j.csda.2017.05.009
3. Bilton, P., Jones, G., Ganesh, S., Haslett, S.: Regression trees for poverty mapping. Australian New Zealand J. Stat. **62**(4), 426–443 (2020). https://doi.org/10.1111/anzs.12312

4. Castagnetto, J.: jmcastagnetto/ubigeo-peru-aumentado: Versión 2.0, corregida y aumentada (2021). https://doi.org/10.5281/zenodo.5248512
5. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2939672.2939785
6. Frink, N., Schmid, T.: Small area estimation with generalized random forests: Estimating poverty rates in Mexico. arXiv preprint (2024). https://arxiv.org/abs/2406.03861, arXiv:2406.03861
7. Frink, N., Schmid, T.: Small area prediction of counts under machine learning-type mixed models. Computational Statistics & Data Analysis, p. 108218 (2025)
8. GeoGPSPerú: Base de Datos Perú Shapefile SHP MINAM. https://www.geogpsperu.com/2014/03/base-de-datos-peru-shapefile-shp-minam.html (2014). Accessed 13 June 2025
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). http://www.deeplearningbook.org
10. Instituto Nacional de Estadística e Informática: REDATAM Census Data Processing 2017. https://censos2017.inei.gob.pe/redatam/ (2017). Accessed July 13, 2025
11. Instituto Nacional de Estadística e Informática: Final Results of the 2017 National Censuses: XII Population, VII Housing, and III Indigenous Communities. INEI, Lima (2018)
12. Instituto Nacional de Estadística e Informática (INEI): Demographic and Family Health Survey - ENDES 2019: National and Departmental Report (2020). https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Endes2019/pdf/Libro.pdf. Accessed 13 July 2025
13. Instituto Nacional de Salud : Situación actual de la anemia. https://anemia.ins.gob.pe/situacion-actual-de-la-anemia-c1. Accessed 13 July 2025
14. Jiang, J., Rao, J.S.: Robust small area estimation: An overview. Annual Rev. Stat. Appl. **7**(1), 337–360 (2020). https://doi.org/10.1146/annurev-statistics-031219-041212
15. Kitaw, B., Asefa, C., Legese, F.: Leveraging machine learning models for anemia severity detection among pregnant women following ANC: Ethiopian context. BMC Public Health **24**, 3500 (2024). https://doi.org/10.1186/s12889-024-21039-x
16. Krennmair, P., Schmid, T.: Flexible domain prediction using mixed effects random forests. J. R. Stat. Soc.: Ser. C: Appl. Stat. **71**(5), 1865–1894 (2022). https://doi.org/10.1111/rssc.12600
17. Nikparvar, B., Thill, J.C.: Machine learning of spatial data. ISPRS Int. J. Geo Inf. **10**(9), 600 (2021). https://doi.org/10.3390/ijgi10090600
18. Siddiqa, M., Shah, G., Butt, M., Kamal, A., Opoku, S.: Early childhood anemia in ghana: prevalence and predictors using machine learning techniques. Children **12**(7), 924 (2025). https://doi.org/10.3390/children12070924
19. Sikov, A., Cerda-Hernandez, J.: Estimating the prevalence of anemia rates among children under five in peruvian districts with a small sample size. Stat. Methods Appl. **32**(5), 1779–1804 (2023). https://doi.org/10.1007/s10260-023-00698-x
20. Sugasawa, S., Kubokawa, T.: Small area estimation with mixed models: a review. Japanese J. Stat. Data Sci. **3**(2), 693–720 (2020). https://doi.org/10.1007/s42081-020-00076-x
21. Tesfaye, S., Seboka, B., Sisay, D.: Application of machine learning methods for predicting childhood anaemia: analysis of Ethiopian Demographic Health Survey of 2016. PLoS ONE **19**(4), e0300172 (2024). https://doi.org/10.1371/journal.pone.0300172

22. Viljanen, M., Meijerink, L., Zwakhals, L., van de Kassteele, J.: A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of netherlands. Int. J. Health Geogr. **21**, 4 (2022). https://doi.org/10.1186/s12942-022-00304-5
23. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992). https://doi.org/10.1016/S0893-6080(05)80023-1
24. Zemariam, A., Yimer, A., Abebe, G., et al.: Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia. Sci. Rep. **14**, 9080 (2024). https://doi.org/10.1038/s41598-024-60027-4

Socioeconomic Factors Associated with Depression in Peru: A Machine Learning-Based Study Using ENDES 2016–2023

Fernando Bendezú-Galarza[1] and Miguel Nunez-del-Prado[1,2(✉)]

[1] Universidad Ricardo Palma, Lima, Peru
miguel.nunez@upc.pe
[2] Universidad Peruana de Ciencias Aplicadas, Lima, Peru

Abstract. Globally, depression has seen a significant increase of 25% in the context of the COVID-19 pandemic. This study investigates the socioeconomic, biological, and social factors associated with depression in Peru. Previous Peruvian research primarily relied on cross-sectional analyses with traditional regression techniques, limiting the ability to detect complex, nonlinear interactions between determinants. Addressing these gaps, this study applies explainable machine learning methods–including Random Forest, Gradient Boosting, and XGBoost–alongside interpretability tools like SHAP to identify and rank key predictors of depression severity over multiple years. Model performance was robust, with XGBoost achieving high accuracy (82%) and AUC-ROC (0.89). These findings underscore the importance of integrated mental and physical health care interventions, social protection, and gender-sensitive strategies tailored to Peru's socio-demographic landscape. The study demonstrates the value of machine learning to provide transparent, actionable insights for targeted public health policies aimed at reducing depression's burden. Its multi-year, interpretable approach advances understanding beyond prior analyses, emphasizing disability and socioeconomic determinants as urgent priorities to improve mental health outcomes in Peru. This research offers significant contributions to evidence-based mental health policy formulation in a middle-income country context.

Keywords: First keyword · Second keyword · Another keyword

1 Introduction

Depression is one of the most prevalent mental health disorders worldwide, affecting more than 280 million people according to the World Health Organization [19]. It contributes substantially to the global burden of disease, causing a loss of approximately 12 billion working days peryear, and generating an estimated

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 495–509, 2026.
https://doi.org/10.1007/978-3-032-20322-9_34

global economic cost of around USD 1 trillion annually due to reduced productivity and disability [20]. Depression is associated with profound negative impacts on individual well-being, social relationships, and economic participation, making it a leading cause of disability worldwide [19].

In Peru, depression represents a serious public health challenge with broad social and economic consequences. Epidemiological evidence suggests that roughly 5% of Peruvians experience clinically relevant depressive symptoms. Depression in Peru has been associated with a lower quality of life, strained family and community relationships, and, in severe cases, suicidal behavior. The financial consequences are also significant, as depression increases healthcare care costs and reduces participation in the workforce. Due to existing inequalities in Peru, vulnerable groups - including women, older adults, and people with chronic diseases - may be especially affected.

Depression is a complex disorder influenced by biological, psychological, and social factors. Among its determinants are chronic diseases such as hypertension and diabetes, unhealthy lifestyle behaviors, weak social support networks, and adverse economic conditions. Gender, age, and regional disparities further modulate the risk of depressive symptoms. For example, higher rates of depression have been reported among Peruvian women, residents of the Andean region, and people with chronic diseases [9], while disability has been shown to substantially increase the risk of severe depression among older adults [3]. However, these investigations mainly relied on cross-sectional data for a single year and traditional regression-based statistical methods.

Previous investigations of depression in Peru have generally relied on cross-sectional data for a single year, with classical regression-based statistical methods to identify associations between variables. Although these approaches have contributed valuable information, they do not fully capture non-linear relationships or complex interactions among multiple risk factors, which can limit their explanatory power and practical application in policymaking.

Moreover, traditional regression models offer limited interpretability about the relative importance of predictors. Policymakers and practitioners increasingly demand transparent and explainable methods that can highlight which variables are most influential in predicting depression levels. These explainable models can help design targeted interventions and allocate resources more effectively in mental health strategies.

This study aims to address these gaps by analyzing data from the Peruvian Demographic and Family Health Survey (ENDES) for the period 2016–2023, applying explainable machine learning techniques to identify and rank the contributions of social, economic, and biological factors to depression severity measured through the PHQ-9. By providing clearer and more interpretable insights into these relationships, this research intends to support evidence-based mental health policies, prioritize preventive actions, and foster cross-sectoral collaboration to mitigate the personal and societal burden of depression in Peru.

2 Theoretical Framework

Depression has been described in the psychiatric literature as a mood disorder characterized by persistent sadness, loss of interest, low self-esteem, guilt, fatigue, poor concentration, and disturbances in sleep and appetite [2,19]. It is categorized under major depressive disorder, dysthymia, substance-induced depression, and medically related depression per the DSM-5 [2], with ICD–10 providing further classification by episode type and severity [18].

Depression is heterogeneous in its presentation. The Diagnostic and Statistical Manual of Mental Disorders (DSM–5) distinguishes major depressive disorder, persistent depressive disorder (dysthymia), substance–induced depression and depression due to other medical conditions [2]. The International Classification of Diseases, 10th Revision (ICD–10) likewise differentiates episodic and recurrent depressive disorders and bipolar affective disorder [18]. Severity can be graded as mild, moderate or severe on the basis of symptom burden and functional impairment [13]. Etiologically, depression arises from an interplay of biological factors (genetic predisposition, neurotransmitter dysregulation, neuroanatomical changes), psychological factors (negative cognitive schemas, unresolved emotional conflicts) and social or environmental factors (stressful life events, lack of social support, poverty) [11,17]. Recognising these multiple pathways is crucial when examining depression within the Peruvian context.

Symptom assessment is commonly conducted using standardized instruments such as the Beck Depression Inventory (BDI), the Center for Epidemiologic Studies Depression Scale (CES-D), and the Hospital Anxiety and Depression Scale (HADS) [4,15,21]. In Peru, the Patient Health Questionnaire-9 (PHQ-9) has been included in the ENDES survey to evaluate depressive symptoms at the population level [13].

The ENDES survey uses the nine–item Patient Health Questionnaire (PHQ–9) to measure depressive symptoms. Each item is scored from 0 ("not at all") to 3 ("nearly every day"), yielding a total score from 0 to 27; scores of 5–9 denote mild symptoms, 10–14 moderate, 15–19 moderately severe and 20 or more severe depression [13]. The PHQ-9's brevity and validated psychometric properties make it suitable for large-scale surveys, although it is not a substitute for clinical diagnosis. Complementary scales such as the Beck Depression Inventory [4], CES-D [15] and Hospital Anxiety and Depression Scale [21] are also widely used in research and clinical settings.

Globally, various risk factors for depression—such as female gender, increasing age, chronic illnesses (e.g., hypertension, diabetes), unhealthy behaviors, low socioeconomic status, and limited social support—have been consistently reported [11,20]. Within the Peruvian context, ENDES-based analyses confirm elevated depression rates among women, Andean-region residents, and individuals with disabilities [3,9].

In recent years, Machine Learning (ML) models—including random forests, neural networks, and ensemble methods—have emerged as promising tools for mental health prediction due to their ability to handle complex, nonlinear relationships [17]. A number of scoping reviews and bibliometric studies have high-

lighted an increasing focus on applying ML to social media and health data to detect depression and anxiety [1,12]. One scoping review reported diagnostic models detecting anxiety and depression via social media posts and activity patterns, primarily using data from platforms like Twitter, Facebook, and Instagram [1]. Another bibliometric analysis documented the growing trend in social media–based ML mental health research, emphasizing both technical advancements and interdisciplinary collaboration [12].

Nonetheless, ML applications in this domain face significant challenges—namely: (a) biases due to demographic imbalances and language differences; (b) methodological inconsistencies, such as inadequate validation, feature selection, and transparency; and (c) privacy and ethical concerns [6]. These issues can limit both the reliability of models and their adoption in health policy.

Explainable AI methods—such as SHAP and LIME—have been introduced to address the interpretability challenge by quantifying feature contributions to model predictions [14,16]. However, their application remains limited in public health surveillance and national survey data analysis.

In summary, while prior Peruvian depression research has relied on traditional statistics and single-year cross-sectional surveys, the integration of explainable machine learning offers the potential to better handle complex relationships, advocate for transparency, and draw more robust, policy-relevant conclusions from multi-year national data.

3 Methodology

3.1 Data and Source

This study uses data from the Peruvian Demographic and Family Health Survey (ENDES), conducted annually by the Instituto Nacional de Estadística e Informática (INEI). ENDES provides nationally representative information on health and demographic indicators from Peruvian households. The data analyzed cover the period from 2016 to 2023, allowing analysis of recent trends with robust sample sizes and consistent survey methodology [10].

3.2 Depression Measurement (PHQ-9)

Depression symptoms in ENDES are measured using the Patient Health Questionnaire-9 (PHQ-9), a validated screening tool extensively used in epidemiological research. PHQ-9 consists of nine items scored from 0 (“not at all”) to 3 (“nearly every day”), generating a total score from 0 to 27, with higher scores indicating greater severity [13].

Consistent with previous analyses of depression in Peru [3,9], this study adopts an applied, correlational design to identify socioeconomic and health-related factors associated with depressive symptoms. The data come from the Encuesta Demográfica y de Salud Familiar (ENDES) conducted by the Instituto Nacional de Estadística e Informática (INEI) [10]. ENDES employs a bi–stage, probabilistic, stratified sampling design at the departmental level with an annual

sample of approximately 36 650 households. Adults aged 18 years or older are surveyed in person by trained interviewers using electronic questionnaires, ensuring nationally representative coverage [10]. Only individuals residing in households are included in the mental-health module.

3.3 Variables and Predictors

The variables analyzed were categorized into social, economic, and biological factors. Social variables included marital status, family size, household type, and region. Economic factors encompassed employment status, education level, and socioeconomic quintile. Biological factors included age, gender, chronic illnesses (e.g., diabetes, hypertension), BMI, and disability status, selected based on previous literature and their consistent presence in ENDES datasets [3,9].

Annual ENDES files (2016–2023) were downloaded and harmonised to create a single analytic dataset. Household and individual identifiers were standardised, and variables were cleaned and recoded. The PHQ-9 total score was treated both as a five-level categorical variable (none, mild, moderate, moderately severe, severe) and as a binary indicator (depressed vs. not depressed). New predictors were engineered to capture lifestyle and household characteristics, including the percentage of days per month consuming healthy foods, the proportion of children under five in the household, a dichotomised marital-status variable and a three-level wealth index. Continuous variables such as age, body-mass index and years of education were normalised, while categorical variables were converted to dummy indicators.

As previous studies have noted, depression is relatively infrequent in community surveys [9]. To mitigate severe class imbalance, the majority "no depression" class was downsampled prior to training. Four supervised models were fitted: Random Forest, logistic regression, XGBoost and a multilayer perceptron (MLP). Hyperparameters—for example, the number of trees, tree depth and learning rate in XGBoost [7], and the penalty and mixture parameters in logistic regression—were tuned via random search and five–fold cross-validation. Each model was coupled with its preprocessing pipeline using workflow objects, and performance was assessed on a held-out test set using accuracy, precision, recall and area under the ROC curve. Interpretation methods such as SHAP values [14] and LIME [16] were applied to quantify feature contributions.

3.4 Machine Learning Models and Interpretability

Random Forest. Random Forest (RF) is an ensemble algorithm consisting of multiple decision trees, each trained on bootstrapped samples of the data. Predictions are obtained through aggregation—typically majority voting for classification—reducing variance and improving generalization. RF models efficiently handle nonlinear relationships, high-dimensional data, and correlated variables [5]. Hyperparameter tuning was performed for parameters such as tree depth, number of estimators, and minimum samples per split.

Gradient Boosting. Gradient Boosting Machines (GBM) sequentially build decision trees, each correcting the errors of previous trees by optimizing a differentiable loss function via gradient descent. GBM excels at capturing complex interactions and nonlinear patterns. Important hyperparameters, such as learning rate, tree depth, and number of estimators, were optimized through cross-validation to balance bias-variance trade-off and improve predictive accuracy [8].

XGBoost. XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting implementation designed for computational efficiency, scalability, and improved predictive performance. It includes regularization techniques to reduce overfitting and uses parallelized computations for faster training [7]. Hyperparameter tuning focused on tree depth, learning rate, regularization parameters (L1 and L2), and subsampling ratios.

Model Evaluation and Validation. Model performance was rigorously evaluated using stratified cross-validation with an 80–20 train-test split. Evaluation metrics included accuracy, precision, recall, specificity, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC). Stratification ensured representative sampling of depression severity categories, improving generalization of model performance metrics.

Interpretability: SHAP. To ensure transparency in interpreting predictions from the ML models, SHapley Additive exPlanations (SHAP) were utilized. SHAP applies cooperative game theory principles to determine each feature's contribution to individual predictions, providing local and global interpretability. Specifically, SHAP calculates a unique additive explanation model that quantifies the average marginal contribution of each predictor to model outputs across the dataset, effectively ranking predictors by importance and visualizing nonlinear associations [14].

3.5 Analytical Pipeline

The analytical pipeline consisted of several sequential steps. Initially, data preprocessing addressed missing data, performed imputation, and ensured variable consistency. Exploratory data analysis (EDA) then guided subsequent feature engineering, helping to identify variable distributions and potential biases. Following preprocessing, ML models (RF, GBM, XGBoost) were trained and optimized through cross-validation. Subsequently, performance evaluation metrics were computed, and SHAP values were generated to interpret feature importance. All analyses were performed using Python, leveraging libraries such as scikit-learn for modeling, pandas for data manipulation, and the SHAP library for interpretability analyses.

4 Results

4.1 Dataset

The final analytic sample included 234,671 respondents from the ENDES survey waves conducted between 2016 and 2023, representing a broad cross-section of Peru's adult population. The average PHQ-9 depression score was 4.3 (SD = 4.1), with a median of 3 points. Approximately 18% of respondents met the threshold for moderate to severe depressive symptoms (PHQ-9 $\geq$ 10), indicating a substantial burden of mental health concerns in the population.

Gender differences were pronounced: women reported significantly higher PHQ-9 scores (mean = 5.1, SD = 4.3) than men (mean = 3.5, SD = 3.7), aligning with global patterns in depression prevalence. An age-related gradient was also evident—older adults (60 years and older) averaged a score of 5.7 (SD = 4.5), while the youngest group (18 to 29 years) averaged 3.6 (SD = 3.5), suggesting that depressive symptoms increase with age.

Regional variations were notable. Participants from the Andean region exhibited the highest average depression scores (mean = 5.2), followed by those from the Amazon/Jungle region (mean = 4.7). Coastal residents had the lowest mean score (4.0). These regional differences may reflect disparities in healthcare access, economic conditions, or cultural attitudes toward mental health.

Socioeconomic status showed a strong inverse association with depression scores. Individuals in the lowest wealth quintile averaged a PHQ-9 score of 6.1 (SD = 4.6), while those in the highest quintile reported a mean of 3.1 (SD = 3.2). Chronic illness and disability were also linked to elevated depression scores. Respondents with at least one chronic disease had an average score of 6.9, and those with a reported disability had the highest observed mean of 7.4. These patterns underscore the substantial mental health burden among medically and socially vulnerable populations in Peru.

4.2 Model Performance and Validation

After balancing the classes and tuning hyperparameters, model performance was compared on the test set. Using precision as the primary metric, the multilayer perceptron achieved the highest precision ($\approx$0.67), followed by logistic regression and Random Forest, while XGBoost trailed slightly. When evaluated by the area under the ROC curve, Random Forest provided the highest discriminative ability ($\approx$0.72), marginally outperforming the other models. These results echo findings from other machine-learning studies in mental-health prediction, where no single algorithm universally dominates [17].

Table 1 provides detailed metrics across the three machine learning models. All models showed good performance, but XGBoost was consistently superior, achieving an accuracy of 82%, sensitivity of 81%, specificity of 83%, F1-score of 81%, and an AUC-ROC of 0.89. Gradient Boosting reached 80% accuracy and an AUC of 0.88, while Random Forest slightly lagged with an accuracy of 78% and an AUC of 0.87.

Table 1. Detailed Performance Metrics of Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	F1-Score	AUC-ROC
Random Forest	0.78	0.78	0.81	0.76	0.87
Gradient Boosting	0.80	0.79	0.82	0.79	0.88
XGBoost	**0.82**	**0.81**	**0.83**	**0.81**	**0.89**

In subgroup analyses, performance metrics were robust across sex, age, and region. For example, for women only, the XGBoost model reached an AUC-ROC of 0.90, slightly higher than the pooled sample, suggesting particularly strong generalizability for female subgroups.

4.3 Feature Importance (SHAP Analysis)

Permutation–based importance measures indicated that personal and health factors were the strongest predictors of depressive symptoms. Sex and age contributed the largest fractions of overall importance, followed by healthy–food consumption, years of education, hypertension and marital status. In contrast, household variables such as wealth index, residence location, type of flooring and vehicle ownership contributed very little to model performance. Similar patterns have been observed in regression analyses, where female gender, older age and chronic disease are consistently associated with higher depression scores [3,9].

The global SHAP feature importance plot (Fig. 1) revealed disability status as the most influential variable in predicting PHQ-9 depression scores. It was followed by chronic illnesses (especially diabetes and hypertension), household wealth quintile, education level, gender, and age.

SHAP values showed that disability contributed more than twice the average marginal SHAP value of the next most influential predictor, chronic illness, which underscores the role of disability as a major risk factor in depression severity. Socioeconomic status was the third strongest driver: individuals in the poorest quintile had significantly higher positive SHAP contributions to their predicted depression risk.

Higher educational attainment showed a negative SHAP contribution, suggesting a protective role in mental health. Women showed consistently positive SHAP values for depression risk across the models, confirming the descriptive patterns. This robust ordering of predictor importance was highly consistent across all three ML algorithms, demonstrating the stability of the findings.

4.4 Additional Analyses and Interpretability

While the previous subsection established the global ranking of predictors across the population, it is also important to understand how these same variables influence individual-level predictions. SHAP values enable such personalized interpretability, revealing which specific features drive the prediction in each case–

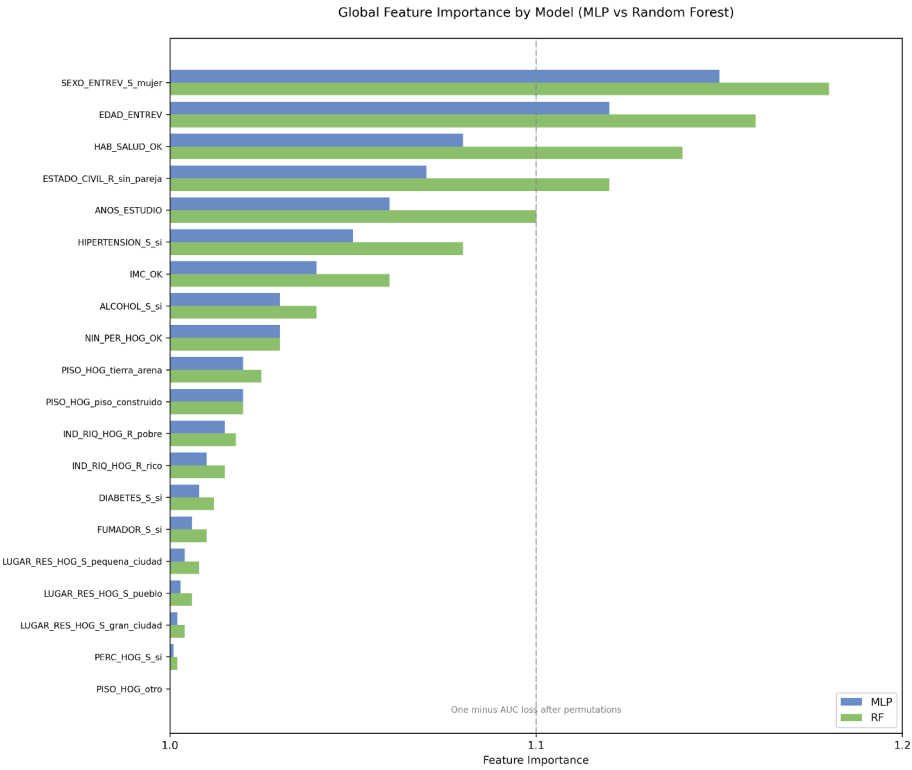


Fig. 1. Global feature importance comparison between MLP and Random Forest models (Loss Ratio).

both in terms of magnitude and direction. To illustrate this, we selected a representative individual from the dataset and computed SHAP values for the top contributing features across both Multilayer Perceptron (MLP) and Random Forest (RF) models.

Table 2 presents the absolute SHAP values for this individual. These values reflect the importance of each feature in shaping the final depression risk prediction, regardless of whether the effect is positive or negative.

However, the magnitude of a SHAP value alone does not indicate whether the feature increased or decreased the predicted score. To address this, Table 3 breaks down the signed contributions of the same variables. For instance, hypertension—despite being the most important feature—had a negative contribution for this individual in both models, meaning it actually reduced the predicted depression score.

This finding illustrates a crucial distinction: some features can be globally important but contextually protective depending on the individual profile. Similarly, female gender, education, and age showed negative contributions, aligning with broader trends while reflecting their personalized interaction with other

Table 2. SHAP Value Magnitudes by Model (Individual 1)

Feature	MLP		Random Forest	
	SHAP Value	Contribution	SHAP Value	Contribution
Hypertension (Yes)	3.306	—	3.306	—
Sex (Female)	0.877	—	0.877	—
Age	0.682	—	0.682	—
Years of Education	−1.230	—	−1.230	—
Marital Status (No Partner)	−0.734	—	−0.734	—
Healthy Habits (OK)	−0.200	—	−0.200	—
Floor: Dirt	−0.653	—	-0.653	—
Floor: Constructed	−1.259	—	−1.758	—
Children in Household	−1.036	—	−1.036	—
Poverty Quintile (Low)	−0.873	—	−0.873	—
Wealth Quintile (High)	−0.573	—	−0.573	—
Alcohol Consumption	0.319	–	0.319	–
Smoker (Yes)	−0.438	–	−0.438	–
Location: Small City	−0.621	–	−0.621	–
Location: Town	−0.589	–	−0.589	–
Location: Large City	−0.357	–	−0.357	–
Vehicle in Household	−0.319	–	−0.319	–
Other Flooring	−0.149	–	−0.149	–
BMI Normal	−0.158	–	−0.158	–
Diabetes (Yes)	−0.182	–	−0.182	–

variables. Minor features such as living in a large city or owning a vehicle had negligible influence, while smoking and alcohol use introduced small but directionally consistent effects.

The "Sign" column indicates the direction of each feature's influence on the predicted PHQ-9 score. A **positive sign (+)** implies that the feature increased the predicted risk of depression for this individual, while a **negative sign (−)** means the feature had a protective (decreasing) effect. A **neutral sign (x)** denotes negligible or no contribution, either because the feature was not active or its impact was too small to register significantly. This symbolic representation helps clarify how each model interprets individual-level predictors beyond their magnitude.

Residual analyses of the main machine learning model indicated no substantial bias across demographic subgroups. The error distribution remained consistent, with no clear evidence of systematic under- or over-prediction in specific population strata. This supports the fairness and generalizability of the predictive models.

Table 3. Signed Contributions by Model (Individual 1)

Feature	MLP		Random Forest	
	SHAP Contribution	Sign	SHAP Contribution	Sign
Hypertension (Yes)	−0.068	−	−0.088	−
Sex (Female)	−0.033	−	−0.047	−
Age	−0.019	−	−0.038	−
Years of Education	−0.013	−	−0.020	−
Marital Status (No Partner)	0.000	x	0.002	+
Healthy Habits (OK)	−0.005	−	0.010	+
Floor: Dirt	0.002	+	0.010	+
Floor: Constructed	−0.001	−	−0.003	−
Children in Household	−0.002	−	−0.015	−
Poverty Quintile (Low)	−0.006	−	−0.003	−
Wealth Quintile (High)	−0.001	−	0.000	x
Alcohol Consumption	−0.002	−	0.001	+
Smoker (Yes)	−0.002	−	−0.002	−
Location: Small City	0.002	+	−0.008	−
Location: Town	0.003	+	−0.005	−
Location: Large City	0.001	+	0.002	+
Vehicle in Household	0.000	x	0.000	x
Other Flooring	−0.001	−	−0.001	−
BMI Normal	−0.001	−	0.001	+
Diabetes (Yes)	0.001	+	0.001	+

Taken together, these complementary analyses reinforce the main findings of the study. They confirm that the relationship between depressive symptoms and factors such as disability, chronic health conditions, economic deprivation, gender, and education is both statistically and substantively significant. Moreover, the use of interpretable machine learning techniques ensures that these associations are not only predictive but also explainable within a transparent analytical framework.

5 Discussion

This study examined the determinants of depression in the Peruvian adult population using nationally representative data from the ENDES surveys conducted between 2016 and 2023, employing explainable machine learning models. The findings identified disability status, chronic illnesses (notably hypertension and diabetes), lower socioeconomic position, limited educational attainment, and

female gender as the most influential predictors of increased depression severity, as measured by PHQ-9 scores.

These associations remained consistent across various model specifications and population subgroups, highlighting the robustness of the results. The presence of chronic or disabling health conditions and structural socioeconomic disadvantages appear to play a central role in shaping mental health outcomes. The convergence of evidence from both descriptive and model-based analyses underscores the importance of targeted mental health policies that consider medical vulnerability, gender disparities, and the social determinants of health.

Compared to earlier work relying on traditional statistical models [3,9], this study provides a broader and more nuanced understanding thanks to the use of explainable machine learning and multi-year data. Previous analyses using logistic regression on ENDES 2017–2018 data had identified similar risk factors, but could not quantify their relative contributions or capture nonlinear patterns. The current results expand upon those findings, confirming the relevance of social determinants and chronic disease, but adding clearer evidence about their relative importance. For example, disability showed nearly double the marginal contribution of the next strongest predictor in SHAP analysis, suggesting it is the most urgent priority for mental health policy intervention.

From a public health perspective, these results highlight multiple areas for targeted prevention and intervention. The strong influence of disability and chronic illnesses on depression severity suggests the need for integrated care models that combine physical and mental health services. Given that socioeconomic disadvantage and low education were also significant risk factors, social protection programs and educational interventions may help mitigate depression rates. Additionally, the persistent gender gap, with women showing higher depression scores, indicates a need for gender-sensitive mental health policies, including screening and tailored treatment.

The convergence of machine-learning and epidemiological evidence suggests that depression in Peru is driven mainly by individual and health-related factors. In particular, being female, of middle or older age, lacking a partner, having low intake of healthy foods, few years of schooling and a diagnosis of hypertension substantially increases the risk of depressive symptoms [3,9]. These findings imply that targeted mental-health interventions should prioritise women over the age of 37, emphasising hypertension management, nutrition education and access to psychological counselling. Because vulnerable women span socioeconomic strata, programmes should incorporate differential fees or subsidies to ensure access for those with limited resources. Integration of mental–health screening into chronic-disease management, as advocated by the World Health Organization [20], could help address the complex interplay between physical and mental health. Routine use of interpretable machine-learning tools may enable policymakers to monitor evolving risk profiles and adjust interventions accordingly.

Importantly, the explainable ML framework provided new capabilities to interpret complex relationships within a high-dimensional dataset. The SHAP

analysis, in particular, made it possible to visualize nonlinear patterns – for instance, the threshold effect of disability and the gradient effect of socioeconomic status – that would be difficult to identify with traditional regression methods. These patterns offer clearer, more actionable information for policymakers and mental health practitioners in Peru.

Nevertheless, the study has some limitations. Although ENDES data are nationally representative and high-quality, they are cross-sectional, which prevents establishing causality between depression and its determinants. There is also potential reporting bias due to self-reported data, which may understate or overstate depression symptoms or chronic conditions. Additionally, some relevant variables—such as social support, life events, or trauma history—are not measured in ENDES and could not be included, potentially omitting important confounders.

Future research should explore longitudinal data sources to establish temporal and causal relationships between social determinants and depression. There is also scope to incorporate additional machine learning methods, such as deep learning on unstructured data, or natural language processing for text-based surveys, to complement structured survey analysis. Finally, intervention studies testing whether addressing the highlighted predictors effectively reduces depression severity would be an essential next step.

Several limitations warrant caution. First, the PHQ-9 instrument, though validated [13], dates from the 1990s; newer scales might capture additional dimensions of depression. Second, the ENDES survey design is cross-sectional, so associations cannot be interpreted causally and temporal dynamics are unobserved [9]. Third, ENDES lacks detailed information on social support and family relationships, factors known to influence mental health [11]. Fourth, methodological changes during the COVID-19 pandemic—such as the use of telephone interviews—may introduce survey mode bias.

6 Conclusion

In this study, we analyzed the socioeconomic, biological, and social determinants of depression in Peru using a comprehensive, nationally representative dataset derived from the Demographic and Family Health Survey (ENDES) spanning the years 2016 to 2023. The principal aim was to identify key factors associated with depression severity in the Peruvian adult population and to quantify their relative importance through advanced explainable machine learning techniques, including Random Forest, Gradient Boosting, and XGBoost, enhanced by interpretability methods such as SHAP values. This multi-year, data-driven approach aimed to provide new, transparent insights capable of supporting evidence-based mental health policies tailored to Peru's complex socio-demographic landscape.

Our results confirm that depression constitutes a substantial public health burden in Peru, affecting nearly one in five adults with moderate to severe symptoms, as measured by the validated PHQ-9 screening tool. The machine learning models achieved strong predictive performance, with XGBoost demonstrating the highest accuracy (82).

The most important finding of this research is the dominant influence of disability status on depression severity, which surpassed other critical predictors such as chronic illnesses (notably hypertension and diabetes), socioeconomic deprivation, limited educational attainment, and female gender. Disability proved to be nearly twice as influential as the next strongest predictor, highlighting it as a crucial focus for mental health interventions. Additionally, socioeconomic factors showed a clear gradient effect, with individuals in the poorest wealth quintiles exhibiting significantly higher depression risk.

Our analysis also revealed important demographic disparities: women and older adults experienced notably higher depression symptom severity, aligning with global and regional epidemiological patterns. The interpretable machine learning framework allowed us to uncover nonlinear relationships and context-dependent effects of various factors, thereby extending beyond the capabilities of previous cross-sectional regression analyses prevalent in prior Peruvian research.

These findings have significant implications for public health policy and intervention design. Effective mitigation of depression in Peru must integrate mental and physical health care strategies that prioritize persons with disabilities and chronic illnesses. Tailored socioeconomic support and educational efforts are essential to alleviate the broader social determinants contributing to mental health disparities. Moreover, gender-sensitive mental health programs that address the elevated risk among women can further enhance intervention effectiveness.

In conclusion, this study demonstrates that disability, chronic illness, socioeconomic disadvantage, low education, and female gender are consistently associated with higher depression severity in Peru. By applying explainable machine learning to a large, multi-year dataset, the study provides new, interpretable evidence to inform public mental health policies. These findings may support more targeted interventions, helping to reduce the burden of depression and improve mental health outcomes for vulnerable populations in Peru.

Future research should therefore (i) evaluate alternative or complementary depression scales in national surveys; (ii) incorporate measures of social support and family functioning; (iii) conduct longitudinal studies to clarify causal pathways; and (iv) explore variables with uncertain effects, such as the presence of young children and body-mass index, to determine their relationship to depression. Qualitative and mixed-methods approaches could also enrich our understanding of gender-specific triggers and protective factors.

References

1. Ahmed, A., et al.: Machine learning models to detect anxiety and depression through social media: a scoping review (2022). https://pubmed.ncbi.nlm.nih.gov/36105318/. Accessed June 2025
2. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. American Psychiatric Association Publishing (2013)
3. Barboza, J.J., Soriano-Moreno, A.N., Copez-Lonzoy, A., Pacheco-Mendoza, J., Toro-Huamanchumo, C.J.: Disability and severe depression among Peruvian older

adults: analysis of the Peru demographic and family health survey, endes 2017. BMC Psychiatry **20**, 1–7 (2020)
4. Beck, A., Ward, C., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Arch. Gen. Psychiatry **4**, 561–571 (1961)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Cao, Y., et al.: Machine learning approaches for mental illness detection on social media: a systematic review of biases and methodological challenges. https://arxiv.org/abs/2410.16204 (2024). Accessed June 2025
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
9. Hernández-Vásquez, A., Vargas-Fernández, R., Bendezu-Quispe, G., Grendas, L.N.: Depression in the Peruvian population and its associated factors: analysis of a national health survey. J. Affect. Disord. **273**, 291–297 (2020)
10. Instituto Nacional de Estadística e Informática (INEI): Encuesta demográfica y de salud familiar (endes) (2024). https://proyectos.inei.gob.pe/endes/. Accessed June 2025
11. Kawachi, I., Berkman, L.: Social ties and mental health. J. Urban Health **78**, 458–467 (2001)
12. Kim, J., Lee, D., Park, E.: Machine learning for mental health in social media: bibliometric study (2021). https://www.jmir.org/2021/3/e24870/. Accessed June 2025
13. Kroenke, K., Spitzer, R., Williams, J.: The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)
14. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural. Inf. Process. Syst. **30**, 4765–4774 (2017)
15. Radloff, L.: The CES-D scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. **1**, 385–401 (1977)
16. Ribeiro, M., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
17. Shatte, A., Hutchinson, D., Teague, S.: Machine learning in mental health: a scoping review of methods and applications. Psychol. Med. **49**(9), 1426–1448 (2019)
18. World Health Organization: International statistical classification of diseases and related health problems, 10th revision (ICD-10) (1992). https://icd.who.int/browse10/2019/en. Accessed June 2025
19. World Health Organization: Depression (2021). https://www.who.int/news-room/fact-sheets/detail/depression
20. World Health Organization: World mental health report: Transforming mental health for all (2022). https://www.who.int/publications/i/item/9789240049338
21. Zigmond, A., Snaith, R.: The hospital anxiety and depression scale. Acta Psychiatr. Scand. **67**, 361–370 (1983)

Conjecture Extraction for Proof Autoformalization

Simon Sorg, Wenda Li, and Soumya Banerjee(✉)

University of Cambridge, Cambridge, UK
{tss52,wl302,sb2333}@cam.ac.uk

Abstract. Autoformalization and ATP have each advanced the mechanization of mathematics, yet the translation of informal proofs into fully formalized counterparts remains an open challenge: especially for interactive theorem provers beyond Isabelle. We introduce conjecture extraction, a novel proof autoformalization pipeline tailored to Lean 4 that decomposes an informal proof into individual lemmas (conjectures), formalizes and proves each in isolation, and then reassembles them to recover the original argument. Unlike prior sketch-based methods, our approach is compatible with end-to-end proof generation models and leverages repeated conjecture refinement to incrementally improve performance. We implement an open-source system that integrates off-the-shelf autoformalization LLMs, automated theorem provers, and an online reinforcement-learning loop to optimize both conjecture generation and proof search. On the MiniF2F benchmark, conjecture extraction achieves an absolute improvement of 11.2% points in pass@1 over the Draft, Sketch, and Prove port (DSP) for Lean 4, demonstrating the efficacy of proof decomposition and recomposition. Our results suggest that conjecture extraction not only bridges a gap in proof autoformalization for Lean but also offers a general framework for scaling formalization efforts across diverse proof assistants. We release our code and models to foster further research in large-scale formalized mathematics.

Keywords: AI for Math · Autoformalization · Automated theorem proving

1 Introduction

Formalization is costly, requiring significant labor by a small group of human experts. To aid mathematicians and enable large-scale formalization, autoformalization has been introduced [19]. Autoformalization translates informal statements into formal versions using Large Language Models (LLMs) [19]. Orthogonal to autoformalization, Automated Theorem Proving (ATP) attempts to prove a formal conjecture. Combinations of autoformalization and ATP focus on large-scale conjecture autoformalization to enable more potential training data for ATP models [20].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 510–521, 2026.
https://doi.org/10.1007/978-3-032-20322-9_35

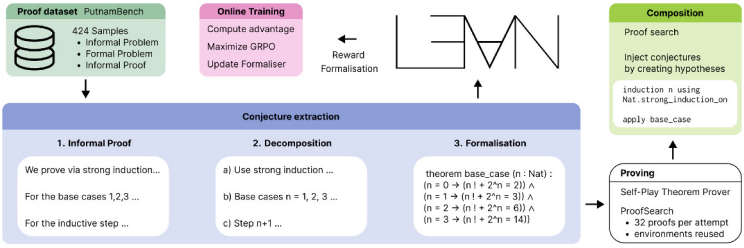


Fig. 1. A general overview of conjecture extraction.

Few works focus on proof autoformalization, translating existing informal proofs into their formalized counterparts [7]. This area is especially relevant for research mathematics, where current ATP solutions still fall short. Works on proof autoformalization target the proof assistant Isabelle [12]. Transferring this to Lean 4 [11] is difficult [10]. The only successful implementation of proof autoformalization is, to the best of our knowledge, a port of *Draft, Sketch, and Prove* (DSP) [7] by [1]. In this work, we propose *conjecture extraction*. We decompose proofs into individual proof steps, prove them individually, and recompose the overarching proof. As our setup does not rely on proof sketching, it is compatible with whole-proof generation models. Further, we improve over DSP for Lean, and show how repeated conjecture extraction further boosts performance.

Our contributions are as follows:

- We present conjecture extraction, a proof autoformalization method designed for Lean 4, compatible with whole-proof generation models.
- We open-source a proof-search implementation combining ATP, autoformalization, and conjecture generation with online reinforcement learning for ATP and autoformalization.
- We provide initial experiments, showing the potential of conjecture extraction, improving absolute 11.2% over DSP in pass@1 proof rate on MiniF2F test [26].

2 Related Work

Our work builds upon existing methods in proof autoformalization, conjecture autoformalization, and ATP, with a focus on the Lean 4 proof assistant [11].

Proof Autoformalization. Proof autoformalization is translating an existing informal proof into a formal one [18]. A prominent method is DSP, where an LLM generates a high-level proof sketch from an informal proof, leaving smaller steps to be solved by pre-existing ATP methods. Originally designed for the Isabelle proof assistant, [10] report problems porting DSP into Lean 4. [1] successfully port DSP into Lean 4. But as LLMs struggle to adhere to consistently adhere to Lean 4 syntax, the performance of DSP in Lean remains limited [1].

Another limitation of DSP is its incompatibility with modern whole-proof generation models [17], as it requires filling in an existing proof structure [7] (Fig. 1).

ATP. LLM-based ATP largely follows two paradigms: proof-step generation and whole-proof generation. Proof-step generation is an interactive process where a model predicts the next tactic based on the current proof state [15], often enhanced with tree-search methods [14]. In contrast, whole-proof generation models produce an entire proof in a single completion [22]. Due to greater computational efficiency, whole-proof generation has become the dominant approach in state-of-the-art ATP systems [17].

Conjecture Generation. The performance of modern ATP models relies on training via expert iteration, a process that requires large-scale datasets of formal conjectures. To overcome the scarcity of human-written formal problems, many works generate data by autoformalizing large corpora of informal mathematics, creating datasets like the Lean workbook. A more scalable approach is to generate entirely new conjectures. The effectiveness of this was demonstrated by [3], who trained a dedicated conjecture generation model and ran expert iteration on its outputs to achieve state-of-the-art ATP performance.

LegoProver. [18] use conjecture generation for proof autoformalization for *LegoProver*, a work closely related to conjecture extraction. LegoProver uses Isabelle, where it aims to solve difficult problems by generating and maintaining a library of proven conjectures. It enhances the DSP framework by using a retrieval mechanism to select useful, already proven premises for a given proof attempt. These premises, along with their proofs, are then copied directly into the context of the LLM, which generates a proof sketch for the main theorem in a two-step process. The informal proof is first rewritten to explicitly state individual proof steps, before this rewritten informal proof is used for a proof-sketch. As LegoProver sketches similar to DSP, the approach is limited to proof-step generation models and does not generalize to whole-proof generation models.

Conjecture extraction also targets individual proof steps but formalizes each step into its own separate conjecture. After these smaller conjectures are proven, the full proof is recomposed. This method only requires the LLM for the natural language task of decomposing the proof, rather than prompting it to generate a complete, syntactically correct formal proof sketch.

Benchmarks. Benchmarks for proof autoformalization are sparse. Only few ATP benchmarks also include informal proofs, enabling proof autoformalization evaluation. A popular ATP benchmark with informal proofs is MiniF2F [26]. PutnamBench [16], a famous ATP benchmark based on the William Lowell Putnam Mathematical Competition, only partially contains informal proofs. For our experiments, we web-scrape informal Putnam proofs of the Internet, leading to 424 samples with informal proofs. Other benchmarks, such as Lean workbook [25], do not provide informal proofs. While ProofNet by [2] contains informal proofs, the Lean 4 port by [22] does not. Therefore, our evaluations focus on MiniF2F and PutnamBench.

3 Methodology

This work's major contribution is *conjecture extraction*, which we cover first. Conjecture extraction requires conjecture autoformalization, and we discuss the two methods we tested next. We lastly end with brief implementation details, focusing on our open-source contribution.

Conjecture Extraction. Decomposing challenging proofs into easier subtasks has been proposed by [18]. Conjecture extraction achieves this by conditioning an LLM on a short prompt together with three problem-specific data points:

- the informal problem statement,
- an informal proof of the problem,
- and a formalized problem statement.

The LLM is tasked with extracting individual isolated proof steps as conjectures from the informal proof. We consider the initial conjecture used to extract conjectures the *parent conjecture*, whose *children* are generated using an LLM. These children adhere to a specified format to facilitate parsing. Specifically, a child conjecture consists of four parts:

1. a preceding *reasoning* block following the chain of thought paradigm
2. a *given* section which introduces variables and their ranges
3. an *assumes* block which states required hypotheses
4. a *shows* area with the conclusion that follows from the variables and assumptions

Note that the differentiation between variables and assumptions is rather blurry. For example, an integer a introduced in the given section could be stated as `positive integer` or as `integer`. In the latter case, positivity could be ensured with an assumption $a > 0$. Both conjectures could yield the same formalization. Still, we decided to separate variables and assumptions, as we found that this pattern arises naturally when prompting various LLMs. We use this format to achieve lower perplexity prompts, which benefits overall generation quality [5]. Once the conjectures are extracted, the LLM output is parsed and autoformalized.

Autoformalization. To formalize conjectures, we test both a general-purpose model and a specialized model trained for autoformalization of competition problems [17].

The existing general-purpose LLMs often generate Lean 3 code, as Lean 4 was released in 2021 and existing Lean code still frequently contains Lean 3 syntax. When zero-shot prompting the generation of Lean 4 code for natural language conjectures, we often faced syntactical errors in the output code containing a mix of Lean 3 and Lean 4 syntax. To mitigate this, we rely on in-context examples [4].

For both models, following [13], we generate multiple formalizations until the first type-checks. Specifically, we query the LLM to generate a formalization, pass

it to the Lean REPL [9], and retry if the formalization is erroneous, as shown in Fig. 2. We retry up to $F = 10$ times, balancing compute and success rates [13]. If no formalization is possible, we skip this natural language conjecture and exclude it from the rest of our pipeline.

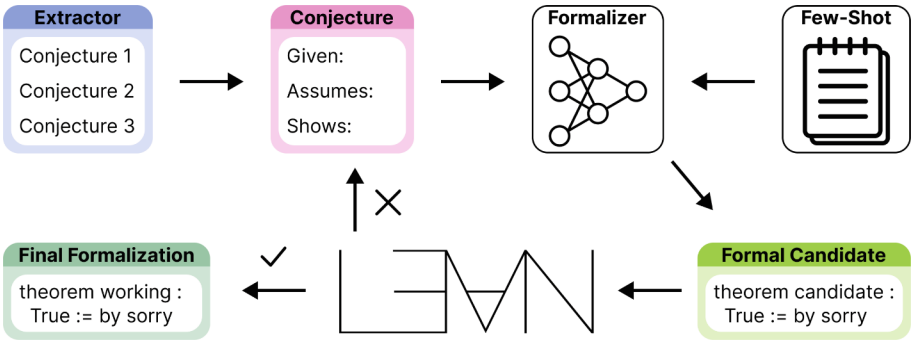


Fig. 2. We formalize each conjecture using one-shot in-context learning [4]. Lean [11] type-checks the conjecture using the Lean REPL [9]. If this fails, we retry the formalization of the same conjecture up to 10 times.

We perform this formalization for each generated conjecture individually, parsing them using the format described above.

Naming Collision. When formalizing many conjectures, we faced the issue of theorem name collisions. Multiple conjectures might receive the same name from our formalization component. If these conjectures are proven and loaded in the same Lean environment, Lean fails, as the name is already in use. To mitigate this, two possibilities arise.

1. Using a separate namespace for each formalized conjecture
2. Introducing a unique affix per theorem that is added to the name

Both variants work equally well and have their implementation challenges. Namespaces are Lean's native way to support the same names in different contexts. We therefore decided to use namespaces to resolve naming collisions. We introduce one namespace per theorem, assigning namespace names via UUID.

Proving. To prove each of the extracted conjectures, we rely on two pre-existing ATP methods. Initial experiments are performed with ReProver Tacgen, an encoder-decoder model for proof-step generation developed by [24]. ReProver supports premises natively, so it does not require the hypothesis injection described below, and thereby serves as a good starting point for our tests.

Later, we generate whole proofs with Self-Play Theorem Prover by [3] instead. For each conjecture, we sample $T = 32$ possible proofs from Self-Play Theorem Prover or $T = 32$ proof steps from ReProver, respectively. For both models, we enhance proof attempts with similar proven conjectures.

Premise Selection. We follow [18] in selecting the most relevant proven conjectures to facilitate further proving. When proving a conjecture, we embed its initial proof state using the ReProver retrieval encoder [24] and retrieve the $k = 10$ nearest premises for it, measured by comparing the cosine similarity. Note that Self-Play Theorem Prover was not trained to deal with additional premises. Simply adding the premises before the theorem to be proven resulted in detrimental performance in our preliminary tests. To nonetheless inject premises, we consider three strategies:

1. Explicitly adding tactics such as `apply premise` and `rw [premise]` for every premise at each proof step. Although straightforward, this method significantly enlarges the search space, impacting efficiency and overall proof rates with the same compute.
2. Converting each premise into a `have` statement embedded within the proof for the overarching statement. This approach enables the LLM used in ATP to select the right premises by adding them to the prompt in a format usable even by whole-proof generation models. However, the resulting proofs are significantly longer, as each proof might contain up to k additional sub-proofs of similar length. Lean only requires the existence of a term of a particular type, rather than the full proof given in the `have` statement.
3. Transforming proven premises into hypotheses by extracting their types and appending the types directly to the theorem statement as hypotheses, which we term *hypothesis injection*. This provides the essential information that such a proof term exists, without exposing unnecessary details.

We choose the third variant as it efficiently provides sufficient information for ATP tools to guide the search space by letting it choose when to apply hypotheses, without unnecessary lengthy proofs.

Internally, hypothesis injection uses the pretty-printed type of a declaration and injects this as Lean syntax in the theorem statement. Over all tests combined, we noticed one edge-case where using the pretty-printed string fails. Definite integrals $\int_a^b f(x)dx$, with the correct syntax in Lean as `∫x in (a)..(b), f x` are pretty-printed as `∫x in a ..b, f x`. Note the missing parentheses and incorrect indentation.

Proof Checking. We execute the generated theorem with its proof in the Lean REPL [9]. The proof-step generation interface is used for ReProver and the whole-proof generation interface for Self-Play Theorem Prover. If unsolved goals remain or an error is encountered, the proof candidate is considered a failure.

We store all proven conjectures in a Lean project and run `lake build` to guarantee the correctness in case of implementation errors in the Lean REPL. To utilize Lean's caching mechanism for faster build times, we place each conjecture in a separate file and import these into the project's main file.

If a conjecture could not be proven, it is re-enqueued for proof search. When next attempted, other new conjectures will have been proven and might be selected in the premise selection step, making the conjecture easier to prove.

Hypothesis Rejection. Similarly to [22], we use hypothesis rejection to filter conjectures with invalid assumptions before attempting a proof. We replace the conclusion of each conjecture with `False` and attempt a proof for `False` using the assumptions. If a proof is found, we discard the conjecture.

Technical Details. We open-source an implementation merging conjecture generation, autoformalization, and ATP via message queueing. It supports both whole-proof generation and proof-step generation models, searching via HyperTreeProofSearch [8], and online reinforcement learning for autoformalization and ATP through the interplay of various components, see Fig. 3. With this, we are the first open-source implementation of online reinforcement learning for ATP and online reinforcement learning for autoformalization. Specific details for individual parts of the implementation can be found in the appendix.

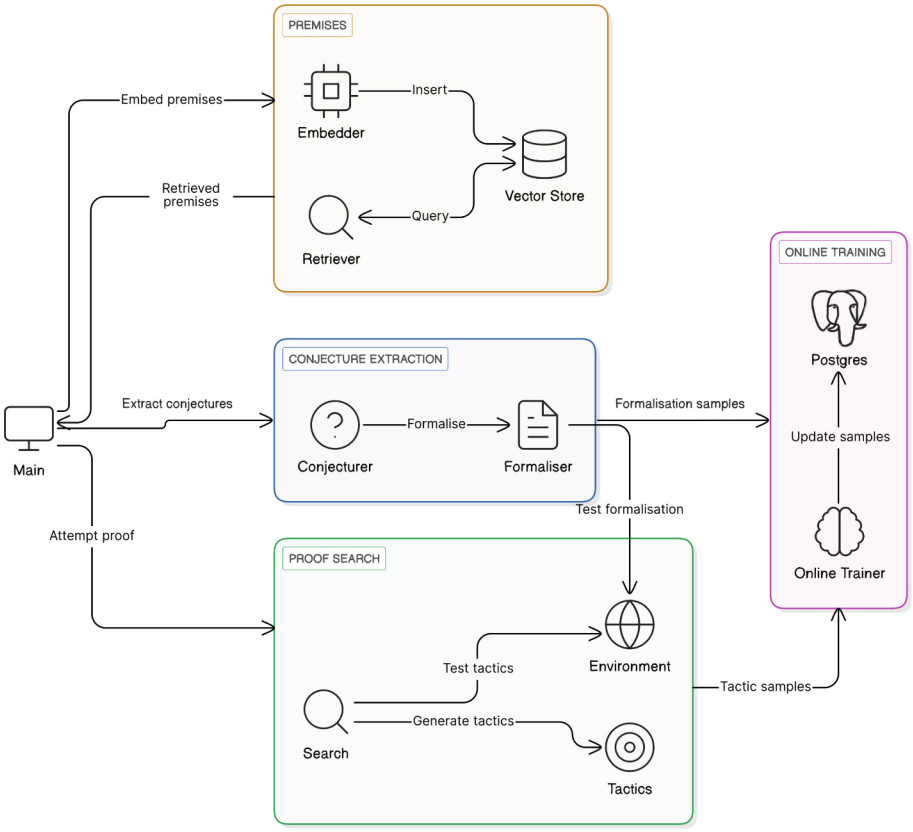


Fig. 3. A simplified overview of the open-sourced implementation.

While we evaluate conjecture extraction, the same implementation can also be used directly for online training of ATP models, without extracting any con-

jectures. Our evaluations show promising results, but should mostly be considered preliminary. We encourage researchers to explore more possibilities of the new pipeline, hoping our contribution will facilitate open-source AI for formalization.

4 Evaluation

As conjecture extraction can be used with any ATP tool, we test it with hammers, proof-step, and whole-proof generation models.

DSP Comparison. Starting the conjecture extraction evaluations, we test it using the same naive hammers as [1]: `aesop`, `simp`, and `linarith`, as the only possible proof steps. For DSP, [1] similarly use GPT-4o [6] to generate one proof sketch per problem. Our setup therefore allows for direct comparison against their results with DSP in Lean.

Conjecture extraction **significantly outperforms DSP** as the current state of the art for proof autoformalization, achieving an **absolute increase of 11.2%** for proof rate on MiniF2F test. DSP errors primarily stem from the LLM not adhering to Lean 4 syntax, resulting in proof-sketch formalization failures. Remarkably, even a naive application of `aesop` to each conjecture without any LLM generation performs similar to DSP (Table 1).

Table 1. Proof rates of DSP, conjecture extraction, and the direct application of `aesop` on the MiniF2F benchmark [26]. We sample one proof sketch and extract conjectures once, respectively.

Method	Validation	Test
Aesop	11.5%	12.7%
Pantograph (DSP)	12.7%	14.7%
Conjecture extraction	**21.3**%	**25.9**%

Subsequently, we compare conjecture extraction with a direct application of the ATP tool for proof-step and whole-proof generation models, discarding DSP due to the low proof sketch success ratio.

Proof-Step Generation. We test proof-step generation using ReProver [24] on PutnamBench [16]. Conjecture extraction using ReProver [24] **solves 2 PutnamBench problems**, **improving over the 0 problems by the default ReProver** [16]. The only better-performing proof-step generation models are InternLM2-StepProver [21] and InternLM2.5-StepProver [20].

ReProver's parameter count is 300 million [23], compared to 7 billion for both InternLM models [21]. We highlight this difference in parameter counts in Fig. 4. Importantly, **all proof-step models outperforming our approach**

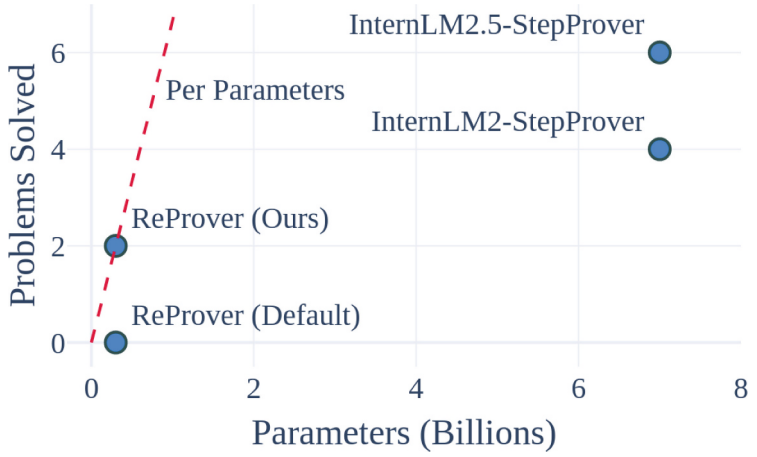


Fig. 4. Proven Putnam problems and parameter counts for proof-step generation models.

are more than 20× **larger**. Beyond proof-step generation, conjecture extraction is the first proof autoformalization method to work on whole-proof generation, which we evaluate next.

Whole-Proof Generation. For whole-proof generation, we test conjecture extraction with Self-Play Theorem Prover [3] on MiniF2F [26]. In our tests, Self-Play Theorem Prover without conjecture extraction proves 135 MiniF2F problems with pass@32. We extract conjectures five times and generate 32 proof candidates for each conjecture, creating a pass@5 × 32 setting.

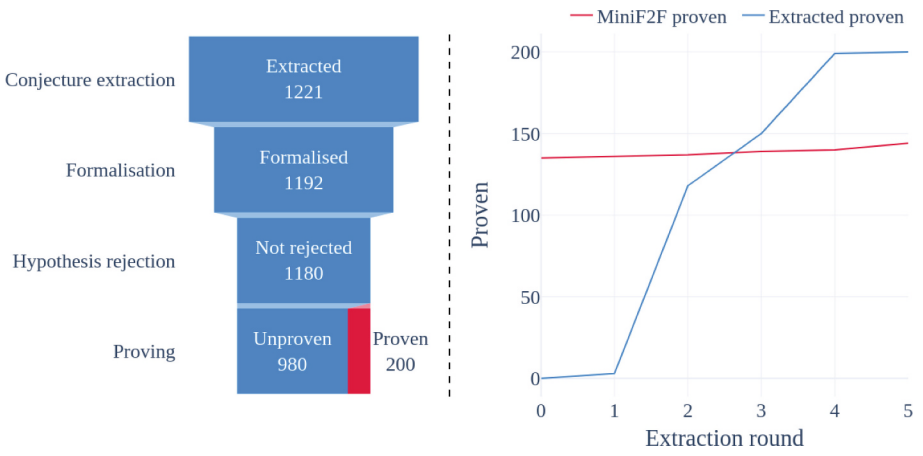


Fig. 5. Left: conjecture counts per pipeline step. Right: more MiniF2F problems are proven with repeated conjecture extraction.

Figure 5 depicts the conjecture counts at the individual stages. 1192 conjectures are extracted and formalized within the five rounds, of which 200 are successfully proven. Twelve conjectures are rejected using hypothesis rejection. After five rounds of conjecture extraction, we cumulatively prove 144 MiniF2F problems. Our results indicate that **repeated conjecture extraction continuously boosts performance** of pre-existing whole-proof generation models, as shown in Fig. 5. We attribute this to different natural language proof steps formed from the same informal proof and different formalizations of similar proof steps, which might be easier to prove.

5 Conclusion

In this work, we have presented *conjecture extraction*, an alternative proof autoformalization method that does not rely on proof sketches. Thereby, conjecture extraction can be used with whole-proof generation models. Our preliminary experiments demonstrate that conjecture extraction substantially outperforms DSP in Lean 4, achieving an 11.2% absolute improvement in the pass@1 proof rate on the MiniF2F test benchmark.

Iteratively extracting and proving conjectures leads to a continuous increase in the number of successfully proven theorems. We open-source an implementation of our pipeline, which combines conjecture generation, autoformalization, and ATP with online reinforcement learning, and can be used for large-scale training and proof autoformalization testing. By providing this framework, we hope to facilitate further research and development in the open-source AI community for formal mathematics. Ultimately, conjecture extraction represents a promising new direction for tackling the challenges of large-scale proof autoformalization, aimed at aiding mathematicians working on formalization projects.

References

1. Aniva, L., Sun, C., Miranda, B., Barrett, C., Koyejo, S.: Pantograph: a machine-to-machine interaction interface for advanced theorem proving, high level reasoning, and data extraction in lean 4. In: Gurfinkel, A., Heule, M. (eds.) Tools and Algorithms for the Construction and Analysis of Systems, pp. 104–123. Springer, Cham (2025)
2. Azerbayev, Z., Piotrowski, B., Avigad, J.: ProofNet: a benchmark for autoformalizing and formally proving undergraduate-level mathematics. In: Poster at the MATH-AI: Toward Human-Level Mathematical Reasoning Workshop, 36$^{\text{th}}$ Conference on Neural Information Processing Systems (NeurIPS 2022) (2022)
3. Dong, K., Ma, T.: Beyond limited data: self-play llm theorem provers with iterative conjecturing and proving. In: The Forty-Second International Conference on Machine Learning (2025)
4. Dong, Q., et al.: A survey on in-context learning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 1107–1128. Association for Computational Linguistics, Miami (2024)

5. Gonen, H., Iyer, S., Blevins, T., Smith, N., Zettlemoyer, L.: Demystifying prompts in language models via perplexity estimation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 10136–10148. Association for Computational Linguistics, Singapore (2023)
6. Hurst, A., et al.: GPT-4o System Card (2024). arXiv:2410.21276 [cs]
7. Jiang, A.Q., et al.: Draft, sketch, and prove: guiding formal theorem provers with informal proofs. In: The Eleventh International Conference on Learning Representations (2023)
8. Lample, G., et al.: HyperTree proof search for neural theorem proving. In: The Thirty-Sixth Annual Conference on Neural Information Processing Systems (2022)
9. LeanProver-Community, T.: A read-eval-print-loop for Lean 4 (2025). https://web.archive.org/web/20250602110502/. https://github.com/leanprover-community/repl
10. Lin, Y., et al.: Goedel-prover: a frontier model for open-source automated theorem proving. CoRR arxiv:2502.07640 (2025)
11. Moura, L.d., Ullrich, S.: The lean 4 theorem prover and programming language. In: Automated Deduction – CADE 28: 28th International Conference on Automated Deduction, Virtual Event, 12–15 July 2021, Proceedings, pp. 625–635. Springer, Heidelberg (2021)
12. Paulson, L.C.: Isabelle, Lecture Notes in Computer Science, vol. 828. Springer, Heidelberg (1994)
13. Poiroux, A., Weiss, G., Kunčak, V., Bosselut, A.: Improving autoformalization using type checking (2025). arXiv:2406.07222 [cs]
14. Polu, S., Han, J.M., Zheng, K., Baksys, M., Babuschkin, I., Sutskever, I.: Formal mathematics statement curriculum learning. In: The Eleventh International Conference on Learning Representations (2023)
15. Polu, S., Sutskever, I.: Generative Language Modeling for Automated Theorem Proving (2020). arXiv:2009.03393 [cs]
16. Tsoukalas, G., et al.: PutnamBench: evaluating neural theorem-provers on the putnam mathematical competition. In: Globerson, A., et al. (eds.) The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
17. Wang, H., et al.: Kimina-prover preview: towards large formal reasoning models with reinforcement learning (2025). arXiv:2504.11354 [cs]
18. Wang, H., et al.: LEGO-prover: neural theorem proving with growing libraries. In: The Twelfth International Conference on Learning Representations (2024)
19. Wu, Y., et al.: Autoformalization with large language models. In: The Thirty-Sixth Annual Conference on Neural Information Processing Systems (2022)
20. Wu, Z., et al.: InternLM2.5-StepProver: advancing automated theorem proving via expert iteration on large-scale lean problems (2024). arXiv:2410.15700 [cs]
21. Wu, Z., Wang, J., Lin, D., Chen, K.: Lean-GitHub: compiling GitHub lean repositories for a versatile Lean prover (2024). arXiv:2407.17227 [cs]
22. Xin, H., et al.: Deepseek-Prover-V1.5: harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search (2024). arXiv:2408.08152
23. Xue, L., et al.: ByT5: towards a token-free future with pre-trained byte-to-byte models. Trans. Assoc. Comput. Linguist. **10**, 291–306 (2022)
24. Yang, K., et al.: LeanDojo: theorem proving with retrieval-augmented language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems, vol. 36, pp. 21573–21612. Curran Associates, Inc. (2023)

25. Ying, H., Wu, Z., Geng, Y., Wang, J., Lin, D., Chen, K.: Lean workbook: a large-scale lean problem set formalized from natural language math problems. In: The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
26. Zheng, K., Han, J.M., Polu, S.: Minif2f: a cross-system benchmark for formal Olympiad-level mathematics. In: The Tenth International Conference on Learning Representations (2022)

Counting Avocados in Orchard Videos Using a YOLO and BoT-SORT-Based Architecture

Diego Albitres Blondet, Franco Galindo Alvarez, and Carlos Fernando Montoya Cubas(✉)

Universidad Peruana De Ciencias Aplicadas, San Isidro, Lima, Peru
{u202015652,u202010807,pcsicmon}@upc.edu.pe

Abstract. Accurate fruit counting is a key component of yield estimation in precision agriculture, enabling the optimization of harvesting, logistics, and export processes. In this work, we propose a system for the automatic counting of avocados in videos captured in real agricultural fields, combining YOLOv8m for fruit detection with the BoT-SORT algorithm for temporal tracking.

The system was validated using a custom dataset composed of public images and original footage collected in commercial Peruvian orchards, under real-world conditions of lighting, occlusion, and high fruit density. The detection model achieved a mAP@0.5 of 0.953, a recall of 0.888, and an F1-score of 0.919. For cumulative counting, the system reached an Absolute Counting Precision (ACP) of 92.68%, along with a Mean Absolute Error (MAE) of 2.0 and a Root Mean Square Error (RMSE) of 3.16, reducing duplicate counts through association and re-identification adjustments.

These results approach the state of the art in controlled environments, such as the work of Du *et al.* (2024), who reported an ACP of 95.33% and an MAE of 3.33. The proximity of our results, despite the more challenging capture conditions, demonstrates that a design focused on cumulative counting enables robust metrics and outcomes applicable to real-world agricultural scenarios, offering a practical solution for producers and companies in the sector.

Keywords: computer vision · precision agriculture · fruit counting · avocado · YOLO · BoT-SORT

1 Introduction

Computer vision techniques are increasingly applied in agriculture, driven by hardware advancements and mature real-time object detection architectures. Automated fruit counting for yield estimation is particularly relevant, optimizing harvesting, logistics, and exports via accurate data [5,7,10,11,13]. Studies show computer vision surpasses traditional methods in accuracy, efficiency, and scalability, with video-based counting being highly promising for field use.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 522–533, 2026.
https://doi.org/10.1007/978-3-032-20322-9_36

However, Peru, a leading Hass avocado exporter [3], lacks validated automatic counting solutions for real production conditions. Current manual counting methods are inaccurate, costly (requiring expert observation), and slow in coverage [8], limiting logistics planning and international competitiveness.

This paper presents a video-based avocado counting model combining YOLO (You Only Look Once) for fruit detection in frames and BoT-SORT for robust object tracking across sequences. This integration enables accurate fruit identification under natural conditions and precise cumulative counting by avoiding duplicates through persistent tracking.

To validate the system, a custom dataset was built using images from commercial Hass avocado orchards and public repositories, capturing natural variations and complexities. The model was trained and evaluated using standard detection and tracking metrics (precision, recall, F1-score, mAP) to balance accuracy and performance.

2 Related Work

Numerous studies have applied YOLO models in agriculture. Osman *et al.* [10] used a custom YOLO for fruit counting via a mobile app. Gremes *et al.* [5] implemented YOLO with tracking for orange counting, while Parr *et al.* [11] used YOLO on depth maps from ToF cameras for grape counting. Other approaches reviewed by Zhai *et al.* [16] include traditional computer vision techniques like HSV color analysis [12] and deep learning methods such as FCNs with hyperspectral imaging for defect detection [15]. Outside agriculture, Tan *et al.* [14] found YOLOv3 outperformed RetinaNet and SSD in speed and precision for pill detection.

A particularly relevant study by Du *et al.* [1] combined an optimized CS-YOLOv5s with enhanced DeepSort for green pepper counting. Their model included attention mechanisms (CBAM), Slim-Neck structures with GSConv, and improved re-identification. While achieving high ACP (95.33%) and low MAE (3.33), their results were obtained under controlled experimental field conditions (stabilized vehicle, segmented lighting). In contrast, our work operates in more challenging, uncontrolled commercial orchard settings. Nevertheless, their approach underscores the value of integrating advanced detection and tracking for issues like occlusion and high density, relevant to our environment. Overall, YOLO architectures are widely favored for their performance and speed across various applications.

3 Methodology

3.1 Model Architecture

The "medium" version of YOLOv8 (YOLOv8m) was used through the library provided by Ultralytics [6], without modifying its internal structure or hyperparameters. The input image size was set to 640×640 pixels, and the batch size

was automatically adjusted based on the available GPU capacity. The training environment was configured using the Google Cloud Colab Enterprise platform, with an n1-standard-2 machine (2 cores, 7.5 GB RAM) [4] and an NVIDIA Tesla T4 accelerator (16 GB VRAM) [9].

Figure 1 shows the overall machine learning pipeline, which includes frame extraction, image segmentation, annotation with Roboflow, and cloud-based training. This pipeline is referenced throughout this section at each corresponding stage.

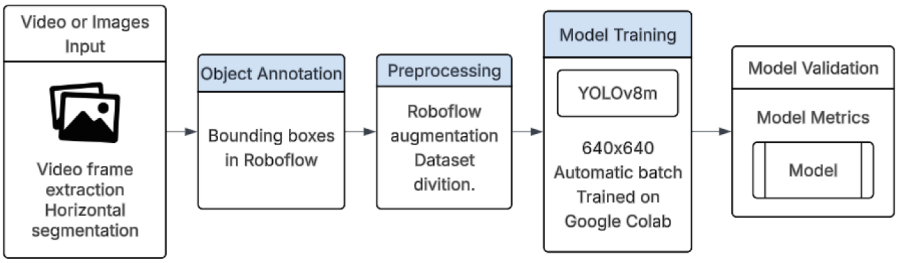


Fig. 1. Machine Learning Pipeline

3.2 Dataset Preparation

The dataset contains a total of 1207 images, of which 124 were obtained from public repositories and 1083 were captured directly from avocado orchards in collaboration with local agribusiness company Frucontrol SAC. These captures, focusing on Hass avocados, included both photos and videos recorded around noon under direct sunlight. The fruits were at a uniform maturity stage, approximately two weeks before harvest. From the videos, one frame per second was extracted for model training. The recordings were made using a Google Pixel 8a at 1080p resolution and 60 fps. The dataset is publicly available at: https://universe.roboflow.com/avocado-workspace/ultimate-avocado/browse.

To ensure the dataset's validity and fairness for real-world application, the collection process specifically targeted challenging conditions. This included high fruit density, significant partial occlusions (with some fruits over 50% obscured by foliage), and high chromatic similarity between the dark green fruits and the surrounding leaves. This focus on occluded instances was intentional, as it presents a more challenging and realistic environment where leaves can be easily confused with avocados. While the fruits were at a consistent maturity level, variations in the camera's distance to the trees provided a diverse range of apparent fruit sizes within the images. The annotation process was conducted using Roboflow (as shown in Fig. 1), with annotators specifically trained to label partially occluded avocados accurately, ensuring high-quality ground truth data for these challenging instances.

As shown in the pipeline in Fig. 1, each capture was split in half to double the number of training samples and increase detail resolution. An example of this division is shown in Fig. 2. Additionally, Roboflow [2] was used to apply data augmentation techniques, generating three variations per training image, resulting in a total of 2897 instances. Transformations included horizontal and vertical flips, ±15° rotations, saturation adjustments (±25%), brightness (±15%), exposure (±10%), blur up to 2.5px, and Gaussian noise in up to 0.1% of the pixels.

Fig. 2. Example of frame division

3.3 Training Process

The dataset was split into three subsets: 80% for training, 10% for validation, and 10% for testing. As illustrated in Fig. 1, the model was trained for a maximum of 100 epochs using YOLOv8m, with early stopping applied if no improvement was observed on the validation metric for 20 consecutive epochs.

3.4 System Operation Architecture

Figure 3 illustrates the overall system architecture for inference. This flowchart details the operational process from video input to the final cumulative count.

The process begins with an input video from the orchard. The video is processed frame by frame. For each frame, the trained YOLOv8m model performs object detection, generating bounding boxes and confidence scores for potential avocados. These detections are then passed to the BoT-SORT algorithm. BoT-SORT handles the core logic of temporal association, matching detections to existing tracks, managing occlusions via its ReID module, and initializing new tracks for newly detected fruits. The system maintains a global list of unique track IDs. The final output is the total number of unique IDs generated, which represents the cumulative count of avocados in the video.

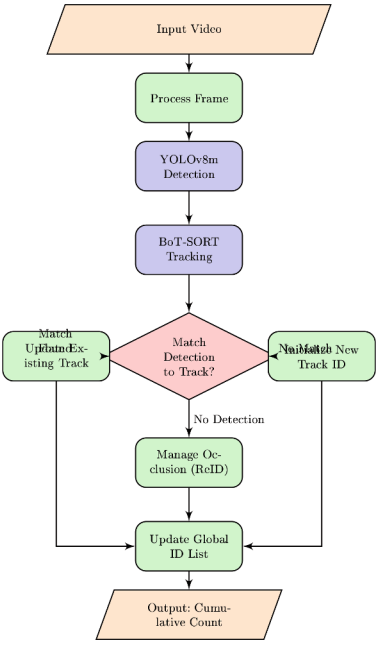


Fig. 3. Flowchart of the overall system operation architecture for inference.

3.5 Object Tracking and Counting

To perform object tracking and cumulative avocado counting, a second pipeline was implemented, as illustrated in Fig. 4, based on the BoT-SORT algorithm. This approach includes an object re-identification module (ReID), which enables more accurate association in scenarios involving partial occlusion or temporary disappearance. The pipeline takes as input the videos previously processed by the YOLOv8m model, which already include detections with bounding boxes and confidence scores. The output consists of annotated videos with persistent IDs for each avocado and the cumulative detection count.

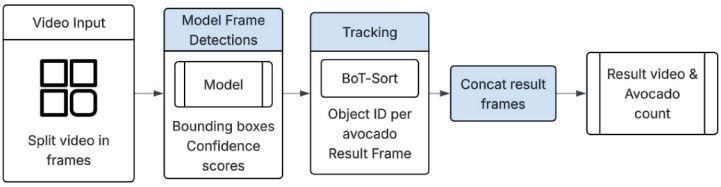


Fig. 4. Tracking Pipeline

The BoT-SORT implementation provided by Ultralytics was used, adapted to integrate directly with the detection outputs of the trained model. Several hyper-

parameters were adjusted to optimize counting accuracy. Specifically, the maximum and minimum association thresholds were set to 0.38 and 0.12, respectively, while the threshold for initializing new tracks was raised to 0.6. For track-to-track matching, a threshold of 0.4 was applied. Proximity and appearance thresholds were set to 0.2 and 0.5, respectively, and the ReID module was enabled.

These parameter settings helped reduce the number of false positives during object association, prioritizing the accuracy of the final count at the expense of lower tracking persistence. The inclusion of the ReID module was particularly important in agricultural environments, where avocados may become partially or fully occluded by foliage. This module enables more reliable visual re-association of previous detections, avoiding the assignment of new, unnecessary identifiers.

This behavior is illustrated in Fig. 5, which presents three consecutive video frames. In the first frame, an avocado is correctly detected and assigned a unique ID. In the second frame, the same avocado becomes temporarily hidden by branches and leaves, resulting in a missed detection. However, in the third frame, once it reappears in the field of view, the system correctly reassigns its original ID. This demonstrates how the system handles temporary occlusions without compromising tracking continuity or duplicating the count, which is essential for accurate fruit estimation.

Fig. 5. Example of temporary occlusion in video. **Frame 1:** the avocado is visible and detected (ID 72). **Frame 2:** the avocado becomes occluded by branches and leaves. **Frame 3:** the avocado reappears and is reassigned its original ID.

3.6 Annotation and Evaluation Data

To validate the accuracy of the counting model, manually annotated videos were required. Three representative videos were selected and labeled using the CVAT.ai platform, identifying all visible avocados in each frame. This allowed for a direct comparison between the system's cumulative count and the ground truth, enabling the calculation of specific tracking and counting evaluation metrics.

4 Experimental Results and Discussion

4.1 Evaluation Metrics

Performance was evaluated using standard metrics for detection, tracking, and counting.

For **detection**, we used Precision, Recall, F1-score, and mean Average Precision (mAP), reported at IoU = 0.5 (mAP@0.5) and averaged over IoU thresholds from 0.5 to 0.95 (mAP@[0.5:0.95]). Precision measures the accuracy of positive detections, Recall measures the proportion of actual fruits found, and F1-score balances the two. mAP provides a comprehensive measure of detection performance across different confidence levels and IoU overlaps.

For **counting and tracking**, following Du *et al.* [1], we used Absolute Counting Precision (ACP), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to assess the final count accuracy against the ground truth (GT). Additionally, Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and ID Switches (IDSW) were used to evaluate frame-by-frame tracking quality, consistency, and spatial accuracy. ACP reflects the relative counting accuracy per video, MAE quantifies the average absolute counting error, and RMSE penalizes larger errors more heavily. MOTA summarizes overall tracking performance, MOTP measures localization precision, and IDSW counts tracking inconsistencies.

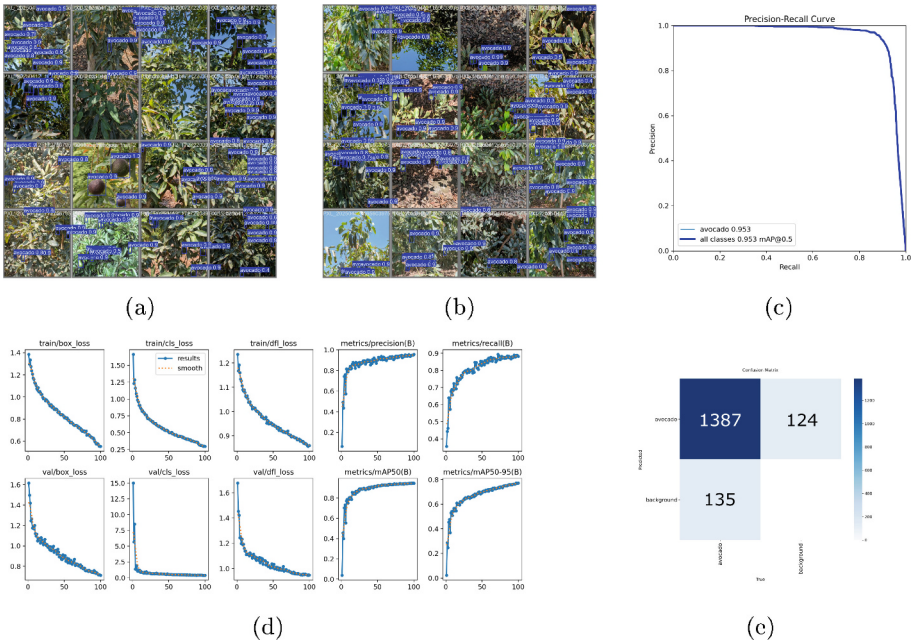
4.2 Detection Experimental Results

On the test set, the YOLOv8m model achieved a precision of 0.953, a recall of 0.888, and an F1-score of 0.919. When evaluated across different Intersection over Union (IoU) thresholds, the model obtained a mAP@0.5 of 0.953 and a mAP@[0.5:0.95] of 0.788. These results demonstrate the robustness of the model in detection tasks, even under varying levels of overlap between predictions and ground truth annotations. Figure 6 presents the main visualizations that quantitatively and qualitatively support the model's performance.

According to the confusion matrix shown in Fig. 6e, out of 1646 avocado instances, the model correctly detected 1387, while 135 false negatives were recorded, mainly due to partial occlusions. Additionally, 124 false positives were identified, mostly corresponding to leaves with a shape similar to avocados.

4.3 Track Experimental Results

In the evaluation of the avocado counting and tracking model using the integration of the YOLOv8m detector with the BoT-SORT tracker, the focus was placed on metrics oriented toward cumulative counting across video sequences. Three key metrics were used to measure counting accuracy: ACP (Absolute Counting Precision), which quantifies the proximity between the predicted and actual (GT) number of avocados as a percentage; MAE (Mean Absolute Error), which calculates the average absolute difference between the predicted and actual



(a) (b) (c) (d) (e)

Fig. 6. (a) and (b) Sample model predictions, (c) Precision-Recall Curve, (d) Cls vs DFL loss (Train vs Validation), (e) Confusion Matrix.

counts; and RMSE (Root Mean Square Error), which measures the dispersion of the counting error, assigning greater penalty to larger deviations.

Tables 1 and 2 present the results per video and the global averages of the aforementioned metrics. In both tables, GT (Ground Truth) denotes the total number of manually annotated avocados in each video, while COUNT indicates the number of avocados automatically detected by the model.

Table 1. Video Count Results

Video	GT	COUNT	ACP (%)	MAE
Video_001	57	57	100.0	0
Video_002	11	13	81.82	2
Video_003	54	60	88.89	6
Video_004	78	78	100.0	0

Table 2 summarizes the average values obtained across the three evaluated videos.

Additionally, Table 3 reports tracking-specific metrics such as MOTA (Multiple Object Tracking Accuracy) and MOTP (Multiple Object Tracking Preci-

Table 2. Global Count Video Results

Algorithm	# Videos	ACP (%)	MAE	RMSE
Yolov8m + BoT-SORT	4	92.68	2	3.16

sion), which reflect the system's frame-by-frame tracking performance. The table also includes the number of ID Switches, corresponding to cases where the same avocado is assigned different IDs throughout the video, potentially affecting the continuity of object tracking.

Table 3. Tracking Metrics Results

Video	MOTA (%)	MOTP (%)	ID Switches
Video_001	67.4	95.1	9
Video_002	67.2	92.4	1
Video_003	43.9	95.7	4
Video_004	39.6	96.2	3

4.4 Discussion

A preliminary version of the detection model, trained solely with public images of avocados, achieved a mAP@0.5 of 0.61 when validated on the current dataset. After incorporating proprietary images captured under real agricultural production conditions, this value improved to a mAP@0.5 of 0.953, representing an increase of over 56%. This progression highlights the impact of dataset diversity and representativeness on the model's ability to detect fruits in open-field commercial scenarios.

Among the reviewed studies, the work by Du *et al.* [1] stands out, as discussed in the Related Work section. Their approach proposes the CS_YOLOv5s architecture, an optimized model based on YOLOv5s. It integrates attention mechanisms (CBAM) and a Slim-Neck structure using GSConv. The dataset used was collected in the experimental field of Hunan Agricultural University, under controlled conditions: recordings were segmented by time of day to control lighting, and a remote-controlled vehicle with mechanical stabilization was employed. In contrast, the proposed system was trained and evaluated using data captured in commercial avocado orchards, without artificial stabilization or environmental control, representing a more challenging and realistic application setting.

To establish a direct comparison with other state-of-the-art architectures, we trained YOLOv7, YOLOv11m, and the CS_YOLOv5s architecture [1] under the same experimental conditions used for our YOLOv8m-based model. The results,

shown in Table 4, demonstrate that our proposed YOLOv8m model (YOLOv8m (Ours)) achieves the highest performance across all key metrics.

Table 4. Comparison of detection metrics between different models

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1-Score
CS_YOLOv5s	0.828	0.557	0.782	0.779	0.780
YOLOv7	0.952	0.729	0.931	0.879	0.904
YOLOv11m	0.946	0.766	0.931	0.889	0.910
YOLOv8m (Ours)	**0.953**	**0.788**	**0.952**	**0.888**	**0.919**

These results reveal two key factors influencing detection quality: the model architecture and the characteristics of the training dataset. In particular, the controlled validation environment used in Du *et al.* [1] promotes more stable detection, but with lower representativeness for unpredictable field conditions typical of commercial avocado production.

Beyond image-level detection performance, a critical component of the proposed system is its ability to perform reliable cumulative counting across video sequences. For this reason, the analysis emphasized specific metrics such as ACP, MAE, and RMSE, which directly reflect the accuracy of the final count–an essential requirement in agricultural tasks that involve precise yield estimation.

The system achieved MOTA values ranging from 39.6% to 67.4%, which are lower than the approximate 90% MOTA reported by Du *et al.* [1]. This discrepancy is attributed to the challenging conditions of the real-world agricultural environment, including temporary occlusions, variable fruit distances, and lighting fluctuations, all of which hinder object visibility and continuous tracking. Nevertheless, despite these limitations, the system attained an average ACP of 92.68%, along with low MAE (2.0) and RMSE (3.16) values, indicating high accuracy in cumulative counting. Moreover, the number of identity switches remained low, reflecting consistent identifier assignment over time.

The tuning of BoT-SORT parameters, such as association thresholds and the activation of the ReID module, proved essential for enhancing the final count accuracy (reflected in ACP, MAE, and RMSE), even at the cost of frame-to-frame tracking continuity (partially explaining the lower MOTA compared to controlled environments). This trade-off was intentional; in agricultural applications for yield estimation, missing occasional detections or track continuity is preferable to generating duplicate counts, thereby prioritizing the reliability of the cumulative result over maximizing frame-by-frame tracking metrics like MOTA.

Overall, both quantitative and qualitative analysis suggest that an effective fruit counting system for agricultural production environments requires not only a robust detection architecture, but also adapted tracking strategies and a representative dataset tailored to the target domain. The integration of these elements

is key to achieving accurate, scalable, and practically useful results under real-world agricultural conditions.

5 Conclusions and Future Work

This work presented an automated system for counting avocados in videos from commercial orchards using YOLOv8m and BoT-SORT. Key findings show the architecture effectively handled real-world challenges (occlusion, variable lighting), achieving high cumulative counting accuracy (ACP 92.68%, MAE 2.0) and strong detection (mAP@0.5 0.953). Comparisons with CS_YOLOv5s [1], YOLOv7, and YOLOv11m validated YOLOv8m's effectiveness and stressed the importance of training on representative data from the target environment (commercial Peruvian plantations). While frame-by-frame tracking (MOTA 39.6%-67.4%) was lower than in controlled settings due to field complexities, consistent tracking (low ID switches, high MOTP 95%) and strategic BoT-SORT tuning prioritized final count accuracy over continuous tracking, minimizing duplications. The results validate the proposed architecture as a precise, viable solution for operational agricultural settings, offering a foundation for future work.

Future work should focus on expanding the dataset with more varied conditions (lighting, occlusion levels, avocado varieties/ripeness) and developing an integrated software solution, possibly incorporating a mobile app or drone-based capture for easier farmer use and enhanced automation. Further exploration into optimizing tracking parameters for different environmental conditions could also improve frame-by-frame consistency without sacrificing counting accuracy.

Acknowledgments. The authors have no competing interests to declare that are relevant to the content of this article.

Disclosure of Interests. This work was carried out in collaboration with the company Frucontrol SAC, which provided essential support for the dataset collection.

References

1. Du, P., Chen, S., Li, X., Hu, W., Lan, N., Lei, X., Xiang, Y.: Green pepper fruits counting based on improved deepsort and optimized yolov5s. Front. Plant Sci. **15** (2024). https://doi.org/10.3389/fpls.2024.1417682
2. Dwyer, B., Nelson, J., Hansen, T.: Roboflow. https://roboflow.com/ (2024), version 1.0
3. Escalante, D., Olivera, J., Miranda, M., Venegas, P.: Peruvian agro-export sector: a competitiveness study on their main products in the period 2010-2019. J. Globalization, Competitiveness Governability **17**, 34–50 (2023). https://doi.org/10.58416/GCG.2023.V17.N2.01
4. Google Cloud: General-purpose machine family for Compute Engine - Compute Engine Documentation (2025). https://cloud.google.com/compute/docs/general-purpose-machines. Accessed 25 May 2025

5. Gremes, M.F., Fermo, I.R., Krummenauer, R., Flores, F.C., Andrade, C.M.G., da Motta Lima, O.C.: System of counting green oranges directly from trees using artificial intelligence. AgriEngineering **5**, 1813–1831 (2023). https://doi.org/10.3390/agriengineering5040111
6. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics (2023), version 8.0.0, AGPL-3.0 License. ORCIDs: 0000-0001-5950-6979, 0000-0002-7603-6750, 0000-0003-3783-7069
7. Liu, Z., Li, X.: An improved yolov5-based apple leaf disease detection method. Sci. Rep. **14**, 17508 (2024). https://doi.org/10.1038/s41598-024-67924-8
8. Maheswari, P., Raja, P., Apolo-Apolo, O.E., Péérez-Ruiz, M.: Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—a review. Front. Plant Sci. **12** - 2021 (2021). https://doi.org/10.3389/fpls.2021.684328
9. NVIDIA: NVIDIA T4 Tensor Core GPU for AI Inference (2025). https://www.nvidia.com/en-us/data-center/tesla-t4/. Accessed 25 May 2025
10. Osman, Y., Dennis, R., Elgazzar, K.: Yield estimation and visualization solution for precision agriculture. Sensors **21** (2021). https://doi.org/10.3390/s21196657
11. Parr, B., Legg, M., Alam, F.: Grape yield estimation with a smartphone's colour and depth cameras using machine learning and computer vision techniques. Comput. Electron. Agriculture **213**, 108174 (2023). https://doi.org/10.1016/j.compag.2023.108174
12. Parraga-Arotinco, E.A., Cárdenas-Martinez, R.P., Misari-Baldeón, J.L., Vilchez-Baca, H.A.: Automated system for sorting blueberries by size and degree of ripeness. In: 2023 International Conference on Control, Robotics and Informatics (ICCRI), pp. 74–80 (2023). https://doi.org/10.1109/ICCRI58865.2023.00021
13. Syazwani, R.W.N., Asraf, H.M., Amin, M.A.M.S., Dalila, K.A.N.: Automated image identification, detection and fruit counting of top-view pineapple crown using machine learning. Alexandria Eng. J. **61**, 1265–1276 (2022). https://doi.org/10.1016/j.aej.2021.06.053
14. Tan, L., Huangfu, T., Wu, L., Chen, W.: Comparison of retinanet, ssd, and yolo v3 for real-time pill identification. BMC Med. Inf. Decision Making **21**, 324 (2021). https://doi.org/10.1186/s12911-021-01691-8
15. Wang, Z., Hu, M., Zhai, G.: Application of deep learning architectures for accurate and rapid detection of internal mechanical damage of blueberry using hyperspectral transmittance data. Sensors **18** (2018). https://doi.org/10.3390/s18041126
16. Zhai, X., Zong, Z., Xuan, K., Zhang, R., Shi, W., Liu, H., Han, Z., Luan, T.: Detection of maturity and counting of blueberry fruits based on attention mechanism and bi-directional feature pyramid network. J. Food Measure. Characterization **18**, 6193–6208 (2024). https://doi.org/10.1007/s11694-024-02640-5

Guidelines for Designing Mobile Applications for Comprehensive Support of Stomach Cancer Patients in Lima: Medical and Spiritual Assistance

Martin Cifuentes Mellado[1], Jean Morin Fuentes[1], Eduardo Díaz[1(✉)], and Carlos Iñiguez-Jarrín[2]

[1] Universidad Peruana de Ciencias Aplicadas, Prolongación Primavera 2390, Lima 15023, Peru
{U202110699,U202115348,pcsijord}@upc.edu.pe
[2] Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador
carlos.iniguez@epn.edu.ec

Abstract. Currently, cancer patients face limitations in accessing technological tools that enable them to comprehensively monitor their illness, receive personalized guidance, and obtain emotional or spiritual support from home. This paper proposes a set of 10 guidelines for the design of a personalized mobile application that provides comprehensive support to patients with stomach cancer. These guidelines are based on a detailed analysis of 31 scientific studies and 10 mobile applications related to the stages of the disease and the challenges faced during treatment. This allows for the implementation of features such as preparation for medical appointments, chemotherapy sessions, adapted nutritional guides, spiritual support, and an AI-assisted chat for specialized medical consultation. This proposal represents an innovation in the field of oncology by integrating emerging technologies with a patient-centered approach. To illustrate the proposed approach, an example of mobile application design using the guidelines is included. Additionally, an experiment was conducted with 10 subjects to measure satisfaction with the web application designed according to the proposed guidelines. It is expected that these guidelines will contribute to the design and development of mobile applications aimed at supporting cancer patients, promoting solutions that improve their quality of life.

Keywords: Mobile application · Stomach cancer · Artificial intelligence · Comprehensive support · Palliative care · Digital health

1 Introduction

Stomach cancer is one of the leading causes of mortality in Peru and ranks among the most prevalent neoplasms, particularly in regions with limited access to specialized oncological services [1, 2]. According to the Global Cancer Observatory, 6,380 new

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 534–549, 2026.
https://doi.org/10.1007/978-3-032-20322-9_37

cases were reported in Peru in 2022, representing 8.8% of all cancer diagnoses nationally [2]. This high incidence, combined with growing treatment costs which increased by 244.9% between 2012 and 2016 according to the Solidarity Health Fund (FISSAL) imposes a considerable economic burden on the healthcare system and affected families [1]. Beyond the medical implications, stomach cancer profoundly impacts patients' emotional, psychological, and spiritual well-being. Research has shown that addressing these dimensions is essential for improving quality of life and treatment adherence, especially in advanced stages [4, 5]. However, most technological interventions and AI-assisted tools developed so far primarily focus on clinical management, often overlooking comprehensive support that includes emotional and spiritual care. In Peru, challenges such as geographic disparities, infrastructural limitations, and budget constraints hinder timely access to quality oncological care, particularly outside Lima and Callao [6]. Furthermore, out-of-pocket health expenditures in the country exceed 28%, exceeding the WHO recommended threshold of 20%, which risks catastrophic financial effects for many households [7]. This context underscores the urgent need for innovative, accessible solutions to support cancer patients in a holistic manner. Mobile health applications powered by artificial intelligence have demonstrated potential to improve treatment adherence, reduce hospital visits, and provide personalized medical recommendations [3]. However, there remains a significant gap in integrating features that attend to patients' emotional and spiritual needs, as well as practical guidance for daily life and treatment preparation.

This paper proposes a set of ten guidelines for designing a personalized mobile application that offers comprehensive support to patients with stomach cancer. These guidelines stem from of 31 scientific literature and 10 mobile applications addressing disease progression, patient challenges, and effective intervention strategies. They encompass thematic focus areas such as spiritual accompaniment and emotional support; design principles including usability, accessibility, and modular architecture; and functional aspects like customized notifications, multimedia educational content, and AI-assisted chatbots for medical inquiries. An illustrative example demonstrates the practical application of these guidelines in a user-friendly mobile platform. Additionally, an experiment was conducted with 10 subjects to measure satisfaction with the web application designed according to the proposed guidelines. This article is structured as follows: Sect. 2 reviews related work on the proposed guidelines. Section 3 presents the definition of the proposed guidelines. Section 4 designs an illustrative example of the mobile application. Section 5 shows early validation. Finally, Sect. 6 shows conclusions, limitations, and future research directions.

2 Related Works

This section presents a Targeted Literature Review (TLR), aimed at identifying scientific studies, best practices, and relevant features that have been applied in the design of mobile applications intended to provide comprehensive support to cancer patients. The search was conducted in renowned bibliographic databases such as Scopus, Web of Science, and IEEE Xplore, using the following search strings: ("cancer patients" AND "mobile application" AND "artificial intelligence" AND "support") and ("spiritual support" AND "AI" AND "cancer").

Inclusion criteria were: (1) scientific studies focused on clinical and emotional support through mobile applications, and (2) the use of artificial intelligence for treatment personalization or monitoring in oncology. Exclusion criteria were: (1) studies with no practical application (purely theoretical in nature), and (2) solutions focused exclusively on diagnosis, without continuous or comprehensive support in the context of cancer. From an initial pool of 165 scientific studies, 10 were selected for detailed analysis. The most relevant studies are grouped and analyzed below according to their area of contribution.

A. Mobile Applications with Artificial Intelligence for Oncological Support

Borsoi et al. [8] developed an AI platform that assists physicians in clinical decision-making for breast cancer treatment. While technically robust, the platform lacks features related to emotional or spiritual support. Michael et al. [9] introduced a digital application for patients undergoing chemotherapy, demonstrating a 30% reduction in unplanned hospital visits. Despite its positive outcomes, the solution does not offer cancer-stage-specific personalization or spiritual support modules. Laymouna et al. [10] evaluated the use of chatbots in the healthcare sector to assess their effectiveness in primary care by providing automated responses to common clinical queries. The study employed a systematic review methodology, analyzing multiple implementations of chatbots based on natural language processing (NLP). The results showed an average accuracy of 79% in answering frequent questions and an improvement in care availability outside clinical hours. However, the authors identified major limitations, including the chatbots' limited capacity for empathy, insufficient contextual personalization, and a lack of adaptation to specific conditions such as gastric cancer. Khambholja and Chhaya [11] highlighted the importance of emotional support for cancer patients and emphasized how AI could be integrated into spiritual support modules. Nevertheless, research on mobile applications incorporating this approach remains scarce.

In summary, although there are multiple initiatives that use artificial intelligence in mobile health applications for oncology, most of them prioritize diagnosis and treatment [1, 8], clinical monitoring [6], or the use of chatbots for emotional support [7, 10], often overlooking a comprehensive approach that also considers the patient's spiritual and emotional well-being.

B. Personalization, Monitoring, and Spiritual Support in Digital Health

Zhang et al. [13] proposed an AI-based application for detecting emotional distress in young cancer patients through voice and language analysis. This approach acknowledges the value of the emotional component but does not incorporate nutritional guidance or spiritual care. Charalambous et al. [14] highlighted in their analysis of European oncology policies that spiritual well-being should be considered a key dimension of cancer treatment. However, they also acknowledged the scarcity of technological tools that integrate this aspect. Hung et al. [15] developed an AI system to recommend clinical trials for patients with head and neck cancer. Despite its technical contribution, the system excludes subjective elements such as the patient's emotional or spiritual state. Finally, Gong et al. [16] explored the use of digital twins in personalized healthcare, emphasizing the importance of data integration for decision-making. Nevertheless, such technologies remain inaccessible in resource-limited contexts like Peru.

The reviewed studies show significant advances in the development of mobile applications targeting cancer patients, including clinical, emotional, and technological functionalities. However, there is a notable lack of proposals that comprehensively address the design of applications specifically for stomach cancer patients, incorporating in a structured way the medical, nutritional, emotional, and spiritual components. This review highlights the need to establish a set of design guidelines that address these dimensions in an integrated manner, constituting the main objective of the present study.

C. Design Guidelines and Usability Considerations for Oncology Mobile Applications

Mescher et al. [17] provide practical guidance for healthcare workers on evaluating and implementing mobile health apps, emphasizing usability, security, and patient-centered design principles critical for effective adoption. Their work underscores the importance of adaptable interfaces and personalized user experiences to meet diverse patient needs. Romero-Ayuso et al. [18] studied the usability of a mobile app aimed at enhancing daily activities for breast cancer survivors. Their findings indicate that intuitive navigation, clear visuals, and modular content improve patient engagement and adherence, elements vital for long-term support applications. Laymouna et al. [10], in their review on healthcare chatbots, also highlight the need for incorporating empathetic design and contextual awareness in AI conversational agents, which remain underdeveloped in oncology applications, limiting their support capabilities beyond basic clinical queries. Keten Edis and Kurtgöz [19] emphasize the role of spirituality in coping with cancer and the spiritual care needs of patients and their families. Their qualitative study suggests that digital tools integrating spiritual support can enhance holistic care but require culturally sensitive and adaptable content. Scherrer et al. [20] proposed a hybrid AI approach for aftercare in cancer patients, integrating clinical data with patient-reported outcomes, which demonstrates the potential for combining medical monitoring with emotional and spiritual well-being features. Caiani et al. [6] highlight regulatory and infrastructural challenges in adopting mobile health technologies, calling for scalable, accessible, and cost-effective solutions tailored to local realities.

Together, these studies reinforce the necessity of design guidelines that transcend pure clinical management, embedding user-friendly design, personalized functionalities, and comprehensive emotional and spiritual support into mobile oncology applications.

3 Definition of Guidelines for Mobile Application Design

This section presents a set of 10 guidelines derived from the literature review and analysis conducted for the design of a mobile application aimed at providing comprehensive medical and spiritual support to stomach cancer patients in Lima. The guidelines are based on the analysis of 31 scientific articles and 10 relevant mobile applications that integrate medical, emotional, and/or spiritual support features for oncology patients. These applications include: CanHope, Belong, PatientMpower, CancerAid, CareZone, My Cancer Coach, Vik, LivingWith, OWise, and War On Cancer.

The scientific articles and applications were analyzed using criteria such as user-centered design, interaction design, spiritual support, psychological accompaniment, and the specific needs of cancer patients within digital health environments. From this review,

common and effective features were extracted and used as a foundation to establish the proposed mobile application design guidelines. Each guideline is identified with the prefix G followed by a number (e.g., G1, G2, G3, etc.). These guidelines are organized into three main categories: (i) Guidelines for the thematic focus of the application; (ii) Guidelines for mobile application design; and (iii) Guidelines on functionality and usability.

A. Guidelines for the Thematic Focus of the Application

G1: Incorporate personalized spiritual support modules. Several scientific studies agree that spiritual support significantly improves the emotional well-being of cancer patients. Delgado-Guay et al. [1] found that spiritual support in patients with advanced cancer is associated with higher quality of life and reduced use of aggressive treatments. Ichihara et al. [2] reported that spiritual interventions reduced emotional pain and increased inner peace among Japanese oncology patients. Similarly, Magharei et al. [3] demonstrated that guided spiritual sessions decrease anxiety and strengthen a sense of purpose. Applications such as CanHope already include resources like meditations, prayers, and faith-based messages personalized according to the user's beliefs. Therefore, it is recommended that the app include a spiritual section with sacred texts, audio materials, and access to religious leaders as part of a more holistic emotional support system.

G2: Implement an AI-based emotional chatbot for personalized psychological support. The use of artificial intelligence (AI) in emotional accompaniment allows for personalized psychological support for oncology patients. García-Prieto et al. [4] concluded that emotional support improves treatment adherence and reduces anxiety. Borsoi et al. [5] employed predictive AI algorithms to analyze symptoms and emotions reported by users, enabling more timely interventions. In a mixed-methods study, Romero-Ayuso et al. [6] observed that digital emotional journals enhance emotional self-awareness and strengthen the patient-caregiver bond. Similarly, Xu et al. [7] validated an AI system using sensors and mobile questionnaires to detect early signs of depression and anxiety. Applications such as Belong provide digital spaces for patients to express their emotions but lack empathetic automated responses. Therefore, it is proposed to include an AI-based emotional chatbot capable of identifying the user's emotional state through natural language processing (NLP), responding empathetically, and encouraging self-expression. This tool would enable continuous, adaptive, and personalized support within the app.

B. Guidelines for Mobile Application Design

G3: Apply user-centered design with a focus on accessibility and intuitive navigation

The application design should follow a user-centered approach, prioritizing information clarity, ease of navigation, and accessibility for people of various ages and levels of digital literacy. Norman and Draper [8] emphasize that usability is a decisive factor in the adoption of health technologies, especially among vulnerable populations. Recommendations include the use of large buttons, intuitive navigation, minimal and legible text, and easily interpretable icons. Cruz-Ramos et al. [9] note that in mHealth applications, it is essential to include visual guides and tutorials for first-time users, as this reduces the abandonment rate. Similarly, Jiménez-Rodríguez et al. [10], in a study

involving older adults, found that adapted design and simple language increased users' confidence and autonomy in using technology. Applications like PatientMpower and OWise incorporate good accessibility practices, such as step-by-step explanatory modules, customizable text sizes, and assisted navigation features that have been positively evaluated by older users. Consequently, it is recommended that the proposed app include a clean interface, adapted to mid-range devices, with options for adjusting text size, intuitive icons, and interactive tutorials to guide patients from their first use. In conclusion, the application should feature easy navigation, large buttons, minimal text, and easily recognizable icons to accommodate users with varying levels of digital literacy. It is also crucial to include visual guides and a permanent help section to support users from their initial interaction with the app. This user-centered design approach enables the creation of an intuitive, accessible, and adaptable interface, significantly increasing the likelihood of adoption and sustained use by oncology patients in diverse contexts

G4: Implement a modular and scalable architecture based on decoupled design patterns. Several studies explore the use of modular and scalable architectures employing design patterns such as Model-View-Controller (MVC) or the Repository pattern to ensure separation of concerns and maintainability. The application should be prepared for future updates, such as adding new support services or integrating with external APIs (e.g., for appointment scheduling or teleconsultations). Obigbesan et al. [1] propose a microservices-based approach for digital health applications where each functional component (such as medical appointments, notifications, or clinical records) is built as an independent module that can be updated or replaced without affecting the entire system. Chimuco et al. [2] emphasize the importance of using clean and decoupled architectures (Clean Architecture) to facilitate integration with external services via API, especially in applications handling sensitive health data. Additionally, Smith and Lee [3] analyzed scalable architectures in health systems and concluded that modular and layered design significantly improves maintainability and enables agile development in resource-limited settings. In conclusion, it is recommended to implement a modular, scalable, and decoupled architecture based on design patterns such as MVC, Repository, microservices, or Clean Architecture. This approach allows the application to adapt to new functionalities without compromising system stability. It also facilitates integration with external services (such as telemedicine platforms or electronic health record systems), enables progressive updates based on user needs, and ensures better long-term technical maintainability, especially in contexts with variable infrastructure or limited resources.

G5: Use warm and neutral color palettes to promote emotional well-being. The visual design of applications for cancer patients must consider the emotional impact that colors can have. Studies such as López-Cobo et al. [13] show that visual environments with warm and soft colors positively influence users' perception of safety and trust—key factors for the acceptance and continued use of a digital tool. In contrast, saturated or alarming colors can cause stress or lead to rejection. Additionally, Singh and Kumar [14] highlight that in healthcare settings, palettes featuring pastel greens, light blues, and beige tones contribute to a calming effect and reduce anxiety among users. Martínez et al. [15] concluded that these colors foster a user-friendly experience, particularly for vulnerable populations such as oncology patients. Successful applications like Life

Manager, CanHope, and PatientMpower use similar palettes, with neutral and pastel tones accompanied by simple, legible iconography. These design choices have been positively rated in usability surveys by patients who value a navigation environment that does not increase their stress or anxiety. In conclusion, it is recommended that the application use a palette of warm, neutral, and low-saturation colors, combined with clear and emotionally neutral iconography, to create a usage environment that conveys calm, trust, and emotional safety—fundamental qualities for patients undergoing cancer treatment.

G6: Ensure compatibility and performance on mid- and low-end devices. Various studies emphasize the importance of mobile health applications being responsive and compatible with a wide range of devices, particularly mid- and low-end Android smartphones, which are predominant in Peru. According to GSMA [16], more than 75% of Peruvians use devices with screens smaller than six point five inches and limited processing capabilities. Therefore, it is essential that the application be optimized for performance, loading times, and include offline functionalities to ensure usability in areas with limited connectivity. Research by Müller et al. [17] and Fernández and Pérez [18] has shown that light-weight mobile apps with low memory usage and offline capabilities increase digital inclusion and sustained use, particularly in rural populations or areas with limited infrastructure. Applications such as CancerAid and LivingWith, used in similar contexts, have successfully implemented these strategies, providing a good user experience on mid- and low-end devices. In conclusion, it is recommended that the mobile application be fully responsive, adaptable to various screen sizes (phones and tablets), optimized for efficient performance on technically limited devices, and include offline functionalities. These conditions are key to ensuring broad adoption and sustained use within the Peruvian context.

C. Guidelines on Functionality and Usability

G7: Implement personalized notifications to reinforce adherence and self-care. Several studies highlight the importance of implementing personalized notification systems that allow users to receive configurable reminders about medication, medical appointments, and motivational messages. These notifications should be adjustable in terms of schedule, frequency, and tone to maximize their effectiveness and user acceptance. Kim et al. [19] demonstrated that personalized reminders significantly increase treatment adherence and enhance perceived usefulness of health applications. Michael et al. [20] reported that intelligent apps with notification systems reduce unplanned visits to healthcare services by 30% and improve therapeutic adherence. Additionally, Lopez et al. [21] found that well-designed notifications with a focus on usability improve patient experience and reduce the feeling of intrusiveness, especially among older adults. Applications such as Belong and PatientMpower integrate configurable and accessible notification systems, which facilitate use among patients with different levels of digital literacy. Therefore, it is recommended that the application incorporate a system of personalized notifications with clear configuration options, prioritizing an intuitive and user-friendly interface for the target group. In conclusion, a system of personalized and programmable notifications should be implemented, allowing users to set alerts related to medication, medical appointments, self-care activities, or motivational messages. These notifications must

adapt to individual preferences in terms of timing, frequency, and content, reinforcing user autonomy and treatment adherence.

G8: Incorporate multimedia educational content to enhance understanding and empowerment. Multiple studies have demonstrated that using educational content in multimedia Multiple studies have demonstrated that using educational content in multimedia formats such as videos, infographics, and audio significantly improves information retention and reduces cognitive overload in cancer patients. According to WHO [22], combining visual and auditory materials facilitates digital health literacy and improves understanding of treatment procedures. In specific investigations, Gómez et al. [23] found that patients who used explanatory videos on procedures and side effects showed greater adherence to treatment and reduced anxiety. Fernández-Rodríguez and Sánchez [24] concluded that auditory and graphic materials tailored to different stages of cancer improve the emotional and cognitive experience of patients.

In the OncoCare application, high-resolution videos segmented by cancer stages explain treatments and nutritional recommendations in a clear and accessible way. The app also integrates interactive infographics and audio materials, allowing users to delve into specific topics based on their interests or needs—features that have been positively rated in usability and patient satisfaction studies. Therefore, it is recommended that the application incorporate varied, accessible, and personalized multimedia content to facilitate understanding and empower patients throughout their oncology treatment journey.

G9: Integrate patient-reported feedback and visual tracking of health indicators. Several studies highlight the importance of including systems that allow patients to regularly record mood, pain levels, energy, and other symptoms, enabling continuous and personalized health monitoring. Zhang et al. [4] developed a digital monitoring system for oncology patients to daily record emotional state and energy levels via a mobile app; this data was then visualized through clear graphs, helping to identify relevant patterns in patient well-being. Similarly, Romero-Ayuso et al. [5] implemented a digital diary using subjective scales to assess pain, emotional state, and functionality, demonstrating through a mixed-methods approach that patients improved their sense of control over their health and communication with healthcare providers. Additionally, Nguyen et al. [6] showed that integrating patient-reported measures with visual dashboards improves patient engagement and facilitates timely clinical interventions in cancer care. Applications like PatientMpower and CancerAid incorporate daily tracking features where users log symptoms and emotions, displaying the data through intuitive graphs and color-coded elements. These tools facilitate self-assessment and provide valuable insights for healthcare professionals to adjust treatments or interventions. Therefore, it is recommended that the application include feedback and visual tracking mechanisms with daily or weekly logs of key variables represented through simple graphs and color-coded elements. This approach supports patient self-management and enhances effective communication with the healthcare team.

G10: Numerous studies emphasize the importance of offering virtual spaces where patients can share experiences, provide advice, and build emotional support networks. These collaborative environments foster empowerment and reduce feelings of isolation.

For example, CancerAid has demonstrated that peer interaction on digital platforms significantly improves the emotional well-being of oncology patients. Smith et al. [7] found that moderated online forums contribute to better psychological outcomes by promoting social connection and information exchange. Similarly, Johnson and Patel [8] highlighted that group chats segmented by cancer type and treatment stage facilitate meaningful peer support while ensuring relevant discussions. Furthermore, Lee et al. [9] showed that combining human moderation with automated monitoring algorithms effectively maintains a respectful, safe, and misinformation-free environment. Therefore, it is recommended that the application include virtual community support spaces such as forums and group chats—moderated by both specialized personnel and monitoring algorithms. This ensures a respectful, safe, and supportive environment that fosters empathy and companionship among users.

Table 1 show a **s**ummary of the ten proposed guidelines.

Table 1 Summary of the ten proposed guides

Category	Guidelines	Details
Guidelines for the Thematic Focus of the Application	G1	Integrate spiritual support with prayers and biblical passages
	G2	Emotional support via AI chatbots
Guidelines for Mobile Application Design	G3	User-centered design with an accessible interface for older adults
	G4	Visual design using warm palettes, neutral tones, and low saturation with clear iconography.
	G5	Responsive, lightweight app optimized for mid- and low-end Android devices with offline capabilities
	G6	Personalized notification system for medication reminders.
Guidelines on Functionality and Usability	G7	Multimedia educational content (videos, infographics, audio)
	G8	Feedback and health status tracking through daily logs.
	G9	Moderated virtual community support spaces (forums, group chats)
	G10	Virtual community

4 Illustrative Example of the Proposed Guidelines

This section presents an illustrative example of the design of a mobile application aimed at patients with stomach cancer, based on the proposed guidelines. The application seeks to offer comprehensive medical and spiritual support through a friendly, personalized, and accessible interface. The following four guidelines are implemented: G1, G2, G3, and G7. Figure 1 shows the implementation of Guideline G1 reflected in the creation of a specific spiritual support module, designed to provide comprehensive emotional support throughout oncology treatment. This section allows the user to access personalized auditory resources such as guided meditations, mindful breathing exercises, positive visualizations, and healing affirmations, all aimed at reducing anxiety, fostering inner peace, and strengthening a sense of purpose. By incorporating these elements according to the patient's beliefs and needs, the application seeks to humanize the recovery process, promoting deeper and sustained emotional well-being. Figure 2 shows the implementation of Guideline G2, focusing on personalized emotional support through artificial intelligence. Here, the user interacts with an emotional chatbot capable of interpreting their mood from the language used in their messages, offering empathetic responses tailored to their emotional state. This feature provides continuous, adaptive psychological support, helping to reduce stress and improve emotional self-awareness.

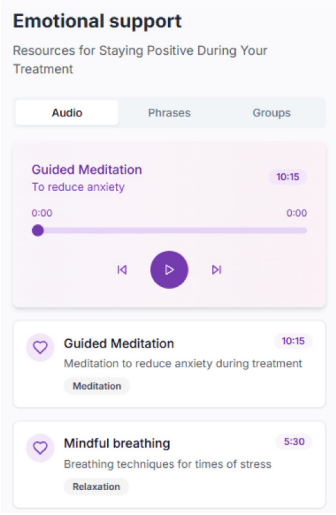


Fig. 1. Spiritual support module with personalized resources for the patient's emotional well-being.

Figure 3 highlights the implementation of Guideline G3, which prioritizes an accessible and user-centered design. The interface features large buttons, intuitive icons, and minimal text, facilitating navigation for users with varying levels of digital literacy. This clear and simple design enhances usability and encourages adoption among diverse patient populations.

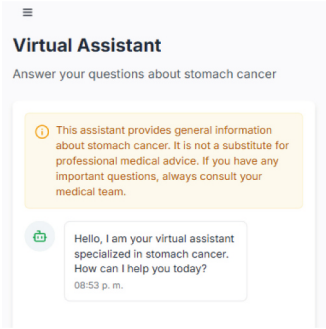


Fig. 2. AI-based emotional chatbot for personalized psychological support.

Figure 4 shows the implementation of Guideline G7, a personalized notification system designed to reinforce treatment adherence and promote self-care. Users can configure reminders for medication, medical appointments, and motivational messages, tailored to their preferences to maximize acceptance and minimize intrusiveness. This system supports patient autonomy and helps optimize clinical outcomes throughout the treatment journey.

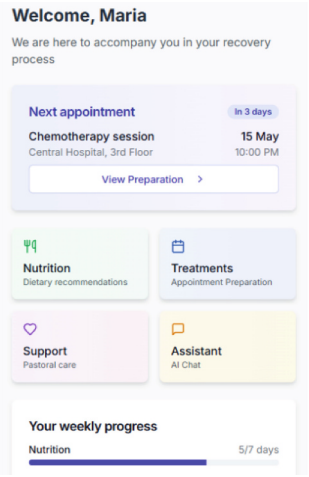


Fig. 3. User-centered design focusing on accessibility and intuitive navigation for oncology patients.

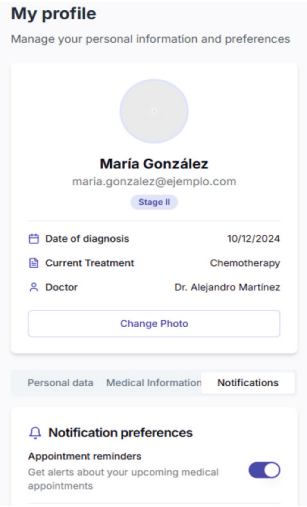


Fig. 4. Implementation of personalized notifications to reinforce adherence and self-care.

5 Early Validation

This section describes an early initial validation (experiment) performed to verify the applicability of the proposed guidelines. The application (see Sect. 4) seeks to offer comprehensive medical and spiritual support through a friendly, personalized, and accessible interface. The validation involved 10 cancer patients from a hospital in Peru who interacted with the application designed according to the proposed guidelines. The patients are referred as subjects. The researchers provided a brief demonstration and task assignments in each section of the application. Subjects interacted with the application throughout the case by selecting from a list of predefined alternatives. For each correct answer, the subject's score increased, and for incorrect answers, they received instant feedback explaining the correct solution. The subjects gave their consent to participate in the experiment. When subjects finished solving the experimental cases, they completed a satisfaction survey (Satisfaction is defined as the contentment and positive attitudes towards the use of a product) [35]. The satisfaction survey was measured online using a 5-point Likert scale questionnaire based on the framework developed by Moody's [36], which defined a framework (based on the work of Lindland et al. [37]) to evaluate satisfaction in terms of Perceived Ease to Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU). This framework has been previously validated and is widely used. The possible answers for each statement in the PEOU, PU, and ITU questionnaire are: Totally disagree, Fairly disagree, Neutral, Fairly agree, and Totally agree. A numerical value is provided to each statement from 1 (Totally disagree) to 5 (Totally agree). Six questions were defined to measure PEOU, the metric was calculated adding the numerical values of the answers and classifying into a rank of five possible values: Rank 1–6: Totally disagree, Rank 7–12: Fairly disagree, Rank 13–18: Neutral, Rank 19–24: Fairly agree, Rank 25–30: Totally agree. For example, if a subject answers 5 questions with Totally agree and 1 questions with Neutral in PU, the result of this metric will be 28 (Totally agree). Six

questions were defined to measure PU; the metric was calculated adding the numerical values of the answers that each subject filled in through the eight questions, the result of this addition is classified into a rank with the five possible options: Rank 1–8: Totally disagree, Rank 9–16: Fairly disagree, Rank 17–24: Neutral, Rank 25–32: Fairly agree, Rank 33–40: Totally agree. Six questions were defined to measure ITU, the metric was calculated adding numerical values of the answers and classifying the result into a rank of two possible values: Rank 1–2: Totally disagree, Rank 3–4: Fairly disagree, Rank 5–6: Neutral, Rank 7–8: Fairly agree, Rank 9–10: Totally agree. Figure 5 shows the satisfaction results.

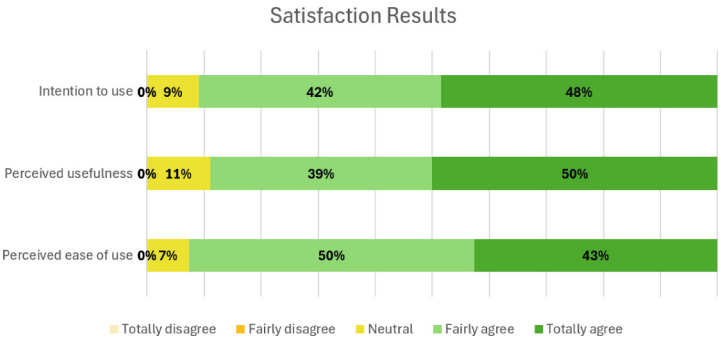


Fig. 5. Satisfaction results

Figure 5 shows a divergent stacked bar of the responses provided by subjects to the questionnaire for this property: For Perceived Ease of Use (PEOU), approximately 43% totally agreed and 50% fairly agreed (over 90% total), suggesting that the majority of subjects considered the application to be easy and potentially useful to use. For Perceived Usefulness (PU), approximately 50% totally agreed and 39% fairly agreed (over 85% total), suggesting that the majority of subjects considered the application to be useful. For Intention to Use (ITU), approximately 48% totally agreed and 42% fairly agreed (over 90% total), suggesting that the majority of subjects intend to use the application. A minority of subjects are undecided and do not perceive the application as useful.

6 Conclusions and Future Work

This paper presents a set of 10 guidelines for designing a personalized mobile application aimed at providing comprehensive support to patients with stomach cancer. These guidelines are based of 31 scientific articles and the analysis of 10 existing conditioning applications for cancer patients. The recommendations are organized into three main categories: (i) thematic focus of the application, which includes spiritual sections, biblical passages, and emotional support; (ii) design and development aspects, such as ease of navigation, large buttons, minimal text, modular and scalable architecture, and the use of warm color palettes; and (iii) functionality and usability features, including customizable and schedulable notifications, multimedia educational content, feedback

and progress tracking, and emotional support spaces. Additionally, an experiment was conducted with 10 subjects to measure satisfaction with the web application designed according to the proposed guidelines. As future work, the following actions are proposed: (i) conduct experimental studies with stomach cancer patients to evaluate the effectiveness of the application implementing these guidelines; (ii) further investigate the specific challenges faced by these patients and explore additional strategies to address their needs in a mobile health context; (iii) analyze and integrate new technological tools that provide emotional and spiritual support, such as interactive relaxation activities and AI-driven chatbots; (iv) establish mechanisms to track patient progress and symptom management over time through the application; and (v) refine and expand the guidelines based on feedback from clinical trials and user experience assessments.

Acknowledgments. "A la Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas por el apoyo brindado para realización de este trabajo de investigación a través del incentivo UPC-EXPOST-2025-2"

References

1. Abuhay, T.M., Robinson, S., Mamuye, A., Kovalchuk, S.V.: Machine learning integrated patient flow simulation: why and how? J. Simul. **17**(5), 580 (2023). https://doi.org/10.1080/17477778.2023.2217334
2. Aminifar, A., Shokri, M., Aminifar, A.: Privacy-preserving edge federated learning for intelligent mobile-health systems. Futur. Gener. Comput. Syst. **161**, 625–637 (2024). https://doi.org/10.1016/j.future.2024.07.035
3. Ayad, N., et al.: Patients' perspectives on the use of artificial intelligence in dentistry: a regional survey. Head and Face Medicine. **19**(1), 23 (2023). https://doi.org/10.1186/s13005-023-00368-z
4. Beglar, N.Y., Rezaei, F., Izadipour, E., TabaTabaei, S.M.: Impact of spiritual end-of-life support on the quality of life for Leukemia patients. Iran. J. Psychiatry. **19**(1), 107–118 (2024). https://doi.org/10.18502/ijps.v19i1.14346
5. Blake, H., Chaplin, W.J., Gupta, A.: The effectiveness of digital interventions for self-management of chronic pain in employment settings: a systematic review. Br. Med. Bull. **151**(1), 36–48 (2024). https://doi.org/10.1093/bmb/ldae007
6. Borsoi, L., Listorti, E., PharmacoEconomics-Open, O.C.: Artificial-intelligence cloud-based platform to support shared decision-making in the Locoregional treatment of breast cancer: protocol for a multidimensional study. Pharmaco Economics-Open. **8**, 945 (2024). https://doi.org/10.1007/s41669-024-00519-1
7. Caiani, E.G., et al.: Standardized assessment of evidence supporting the adoption of mobile health solutions: a clinical consensus statement of the ESC regulatory affairs committee. Eur. Heart J. Digit. Health. **5**(5), 509–523 (2024). https://doi.org/10.1093/ehjdh/ztae042
8. Charalambous, A., Price, R., Jha, P.: Accelerating progress on EU cancer control. Lancet Oncol. **25**(2), 158–160 (2024). https://doi.org/10.1016/S1470-2045(24)00002-0
9. Chimuco, F.T., Sequeiros, J.B.F., Simões, T.M.C., Freire, M.M., Inácio, P.R.M.: Expediting the design and development of secure cloud-based mobile apps. Int. J. Inf. Secur. **23**(4), 3043–3064 (2024). https://doi.org/10.1007/s10207-024-00880-6
10. Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S., Barteit, S.: Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. npj Digit. Med. **5**(1), 162 (2022). https://doi.org/10.1038/s41746-022-00700-y

11. Ichihara, K., Nishiyama, C., Kiyohara, K., Morita, T., Tamura, K.: Nursing Care for Spiritual Pain in terminal cancer patients: a non-randomized controlled trial. J. Pain Symptom Manag. **67**(2), 126–137 (2024). https://doi.org/10.1016/j.jpainsymman.2023.10.016
12. Janani, S.R., Subramanian, R., Karthik, S., Vimalarani, C.: Healthcare monitoring using machine learning based data analytics. Int. J. Comput. Commun. Control. **18**(1) (2023). https://doi.org/10.15837/ijccc.2023.1.4973
13. Keten Edis, E., Kurtgöz, A.: The role of spirituality for coping with cancer and the spiritual care needs of women with breast cancer and their family caregivers in Turkey: a qualitative study. J. Relig. Health. **63**(2), 1475 (2024). https://doi.org/10.1007/s10943-023-01984-4
14. Khambholja, K., Chhaya, V.: HPR95 public-private partnership model in oncology: business opportunities for next generation contract research organizations in the context of CancerX. Value Health. **26**(12), S270 (2023). https://doi.org/10.1016/J.JVAL.2023.09.1413
15. Krones, F., Marikkar, U., Parsons, G., Szmul, A., Mahdi, A.: Review of multimodal machine learning approaches in healthcare. Inf. Fusion. **114**, 102690 (2025). https://doi.org/10.1016/j.inffus.2024.102690
16. Laymouna, M., Ma, Y., Lessard, D., Schuster, T., Engler, K., Lebouché, B.: Roles, users, benefits, and limitations of Chatbots in health care: rapid review. J. Med. Internet Res. **26** (2024). https://doi.org/10.2196/56930
17. Lee, M., et al.: Efficacy of the PRO-CTCAE mobile application for improving patient participation in symptom management during cancer treatment: a randomized controlled trial. Support. Care Cancer. **31**(6), 321 (2023). https://doi.org/10.1007/s00520-023-07779-3
18. Michael, A., et al.: 1835P multi-Centre, randomised controlled trial of digital health cancer solution for cancer patients receiving chemotherapy. Ann. Oncol. **35**, S1085 (2024). https://doi.org/10.1016/J.ANNONC.2024.08.1930
19. Ministerio de Salud: Análisis de la situación del Cáncer en el Perú, 2018 (2018). http://www.dge.gob.pe
20. Mbunge, E., Batani, J.: Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: emerging opportunities, trends and implications. Telematics Inform. Rep. **11**, 100097 (2023). https://doi.org/10.1016/j.teler.2023.100097
21. Morgenstern, J.D., Rosella, L.C., Daley, M.J., Goel, V., Schünemann, H.J., Piggott, T.: "AI'S gonna have an impact on everything in society, so it has to have an impact on public health": a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. BMC Public Health. **21**(1), 40 (2021). https://doi.org/10.1186/s12889-020-10030-x
22. Muñoz, A.: Real-Time Diagnosis Algorithms in Biomedical Applications and Decision Support Tools (2024). https://doi.org/10.3390/books978-3-7258-1300-1
23. Nwankwo, E.I., Emeihe, E.V., Dayo Ajegbile, M., Olaboye, J.A., Maha, C.C.: Integrating telemedicine and AI to improve healthcare access in rural settings. Int. J. Life Sci. Res. Arch. **2024**(01), 59–077 (2024). https://doi.org/10.53771/ijlsra.2024.7.1.0061
24. Obigbesan, O., Graham, K., Benzies, K.M.: Software testing of eHealth interventions: existing practices and the future of an iterative strategy. *JMIR*. Nursing. **7**(1), e56585 (2024). https://doi.org/10.2196/56585
25. Park, M., Choi, E.K., Lyu, C.J., Han, J.W., Hahn, S.M.: Family resilience factors affecting family adaptation of children with cancer: a cross-sectional study. Eur. J. Oncol. Nurs. **56**, 102078 (2022). https://doi.org/10.1016/J.EJON.2021.102078
26. Petersson, L., et al.: Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv. Res. **22**(1), 850 (2022). https://doi.org/10.1186/s12913-022-08215-8

27. Prasongsook, N., Seetalarom, K., Saichaemchan, S., Udomdamrongkul, K.: A pilot study of using smartphone application vs. routine follow-up for patient Care in Advanced non-Small Cell Lung Cancer during the COVID-19 pandemic era. Front. Med. Technol. **4**, 900172 (2022). https://doi.org/10.3389/fmedt.2022.900172
28. Purcell, W.M., Burrell, D.N.: Dynamic evaluation approaches to telehealth technologies and artificial intelligence (AI) telemedicine applications in healthcare and biotechnology organizations. Merits. **3**(4), 700–721 (2023). https://doi.org/10.3390/MERITS3040042
29. Romero-Ayuso, D., et al.: Usability of a mobile phone application to enhance activities of daily living in occupational therapy services for breast cancer survivors. Hong Kong J. Occup. Ther. **36**(2), 128 (2023). https://doi.org/10.1177/15691861231206489
30. Rosenberg, A.R., Weaver, M.S., Fry, A., Wiener, L.: Exploring the impact of the coronavirus pandemic on Pediatric palliative care clinician personal and professional Well-being: a qualitative analysis of U.S. survey data. J. Pain Symptom Manag. **61**(4), 805–811 (2021). https://doi.org/10.1016/j.jpainsymman.2020.09.037
31. Scherrer, A., et al.: A hybrid artificial intelligence solution approach to aftercare for cancer patients. Neural Comput. & Applic. **35**(29), 21381 (2023). https://doi.org/10.1007/s00521-023-08765-w
32. Senthilkumar, T., Arumugam, T., Pandurangan, H., Panjaiyan, K.: Adoption of artificial intelligence in health care: a nursing perspective | Adopción de la Inteligencia artificial en la Atención sanitaria: Una Perspectiva Enfermera. Salud, Ciencia y Tecnologia. **3**, 510 (2023). https://doi.org/10.56294/saludcyt2023510
33. Xu, L., Sanders, L., Li, K., Chow, J.C.L.: Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer. **7**(4), e27850 (2021). https://doi.org/10.2196/27850
34. Zhang, A., et al.: Evaluating the feasibility and acceptability of an artificial-intelligence-enabled and speech-based distress screening Mobile app for adolescents and young adults diagnosed with cancer: a study protocol. Cancer. **14**(4), 914 (2022). https://doi.org/10.3390/cancers14040914
35. S. Association: Systems and software engineering—vocabulary ISO/IEC/IEEE 24765: 2010. Iso/Iec/Ieee. **24765**, 1–418 (2010)
36. Moody, D. L.: The method evaluation model: a theoretical model for validating information systems design methods (2003). https://aisel.aisnet.org/ecis2003/79
37. Lindland, O.I., Sindre, G., Solvberg, A.: Understanding quality in conceptual modeling. IEEE Softw. **11**(2), 42–49 (1994). https://doi.org/10.1109/52.268955

Author Index

A
Acosta Mantaro, Juan Eduardo 419
Aguilar-Encarnacion, Pedro 342
Albujasim, Zainab 314
Alvarez, Franco Galindo 522
Alvarez-Mamani, Edwin 93, 450
Alzamora, Guina Sotomayor 207
Apaza, Honorio 137
Apaza-Cutipa, Renzo 433
Arroyo-Paz, Antonio 125, 358
Azabache Morán, Carlos Alberto 419

B
Balakrishnan, Dhananjay 109
Banerjee, Soumya 510
Barrios, Jhon Edilberto Monroy 300
Bellver-Soler, Jaime 13
Benavides, Dylan 27
Bendezú-Galarza, Fernando 495
Blondet, Diego Albitres 522

C
Calcina-Puma, Esteven 358
Cárdenas-Martínez, Franco 125
Ccansaya-Soncco, Rebeca Araceli 450
Ceh-Varela, Edgar 468
Checcllo, Amy 45
Choque-Quispe, Jadyra Ch'aska 450
Cifuentes Mellado, Martin 534
Coaguila, Maribel E. 137
Collanqui, Elmer A. 137
Condori-Alejo, Henry Ivan 207
Copara, Jenny 280
Córdova, Arcenio 403
Cubas, Carlos Fernando Montoya 522
Cuenca, Erick 403
Cuenca, Víctor 403
Cusi, Enrique Miguel Paco 300
Cusi, Ibeth J. del Pilar 93
Cutipa-Santi, Michael Newton 433
Cutipa-Santi, Yeni 433

D
Díaz, Eduardo 1, 534
D'Haro, Luis Fernando 13
Diallo, Gayo 79, 246
Diaz, David Aragon 13
Díaz, Eduardo 468
Domínguez, Martín 60
Drancé, Martin 246
Drekshagen, Svea 109

E
Eguiluz, Jeymi Melanie Valdivia 300
Esina, Svetlana 387
Espezúa, Soledad 45

F
Falquet, Gilles 280
Farromeque, Eva 419
Fiuri, Ariel 60
Flores, Anibal 137
Francis, Basurco Zapata Gerald 300

G
Gamero, Claudia Lucia Rozas 468
Giraldo, Hugo 419
Gómez Medina, Helen Magdalena 403
Gourdon, Juliette 263
Guragain, Anmol 13
Gutierrez-Pachas, Daniel A. 480

H
Huamán Curo, Edwin 167
Huaman, Elwin 184
Huaman, Jorge Luis 184
Huaman, Wendi 184
Huapaya, Orlando Arturo Roca 1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. A. Lossio-Ventura et al. (Eds.): SIMBig 2025, CCIS 2895, pp. 551–552, 2026.
https://doi.org/10.1007/978-3-032-20322-9

Huarcaya Farfan, Dylan J. 194
Huincho, Miguel Angel Portella 300

I
Iñiguez-Jarrín, Carlos 534
Ibañez-Quispe, Vladimiro 433
Iñiguez-Jarrín, Carlos 342
Inkpen, Diana 314, 387

J
Jallo-Paccaya, Yasumy Maricely 450
Jean-Alphonse, Frederic 263
Jerez Calero, María Elena 403

L
Lazo Vasquez, Ricardo 374
Li, Wenda 510
Limaylla-Lunarejo, María Isabel 222
Lin, Long 13

M
Maita-Vargas, Marcelino 194
Mamani, Anthony 480
Mamani, Lizeth C. 93
Mamani-Lacuta, Deyvis 207
Mancusi Barreda, Giorgio 327
Miranda, Oscar Heredia 468
Morin Fuentes, Jean 534
Moscoso Cuaresma, Julio Ricardo 419
Mougin, Fleur 246

N
Naderi, Nona 280
Navarro Mendoza, Aaron C. A. 194
Nunez-del-Prado, Miguel 109, 419, 495

O
Ochoa-Luna, José 374
Opazo, Juan 234
Osenda, Omar 60

P
Pineda Ancco, Ferdinand 167

Q
Quelca-Velasquez, Daisy Crishyely 433
Quispe, Ninfa 184
Quispe-Mamani, Thalia 450

R
Ramirez, Renzo Manuel Ramos 1
Ramos-Varela, Samuel 13
Razzaq, Misbah 263
Ríos, Alexia 45
Rodríguez, Oldemar 27
Romero, Pamela 263
Rosales, Fiorella Jara 300
Rosero, Luis 480

S
Sahu, Vedant 109
Sandobalín, Julio 342
Sanjinez, Alexandra 45
Santana, Sofía Perón 60
Santisteban, Julio 154
Solaligue, Christian Aranibar 154
Solís, Maikol 27
Sorg, Simon 510

T
Tariverdi, Mersedeh 109
Teodoro, Douglas 280
Ticse, Adriana 419
Ticse, Grycel 419

U
Urbina, Myriam 403
Usurin Arias, Fernando A. 194

V
Velasquez Borasino, Christian 327
Vera-Olivera, Harley 93, 450
Vilchez Marin, Rody 327
Villanueva, Edwin 234
Villegas Suarez, Ariana M. 194

Y
Yvinec, Romain 263

Z
Zemmari, Akka 246
Zuñiga-Rojas, Gabriela 93