Osvaldo Gervasi · Beniamino Murgante ·
Chiara Garau · Ana Maria A. C. Rocha ·
M. Noelia Faginas Lago · David Taniar ·
Yeliz Karaca (Eds.)

LNCS 16769

Computational Science and Its Applications – ICCSA 2026

26th International Conference
Braga, Portugal, June 30 – July 3, 2026
Proceedings, Part III

3 Part III

Springer

Lecture Notes in Computer Science **16769**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Osvaldo Gervasi · Beniamino Murgante ·
Chiara Garau · Ana Maria A. C. Rocha ·
M. Noelia Faginas Lago · David Taniar ·
Yeliz Karaca
Editors

Computational Science and Its Applications – ICCSA 2026

26th International Conference
Braga, Portugal, June 30 – July 3, 2026
Proceedings, Part III

Springer

Editors
Osvaldo Gervasi
University of Perugia
Perugia, Italy

Chiara Garau
University of Cagliari
Cagliari, Italy

M. Noelia Faginas Lago
University of Perugia
Perugia, Italy

Yeliz Karaca
University of Massachusetts
Amherst, MA, USA

Beniamino Murgante
University of Basilicata
Potenza, Italy

Ana Maria A. C. Rocha
University of Minho
Braga, Portugal

David Taniar
Monash University
Melbourne, VIC, Australia

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-032-30493-3 ISBN 978-3-032-30494-0 (eBook)
https://doi.org/10.1007/978-3-032-30494-0

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2027

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

These 3 volumes (LNCS volumes 16767–16769) comprise the peer-reviewed papers from the 6 tracks of the 2026 International Conference on Computational Science and Its Applications (ICCSA 2026), held between June 30 and July 3, 2026, in Braga (Portugal). The peer-reviewed papers from the 57 Workshops are published in a separate set of 9 volumes (LNCS 16758–16766).

The conference was held in a hybrid format, with the vast majority of participants attending in person, hosted by the University of Minho in Braga, Portugal. We enabled virtual participation for those who did not attend the event in person due to logistical, political, or economic constraints by adopting a technological infrastructure based on open-source software (Jitsi + Riot) and a commercial Cloud infrastructure. ICCSA 2026 was organized by the University of Minho (Portugal), the University of Perugia (Italy), the University of Basilicata (Italy), the University of Cagliari (Italy), Monash University (Australia), and Kyushu Sangyo University (Japan).

With the 2026 edition, ICCSA celebrated its 26th anniversary, marking a new chapter in its longstanding history after its memorable 25th anniversary celebrated in Istanbul, Türkiye. Hosted in Braga, Portugal, the conference found a particularly meaningful setting in one of Europe's oldest and most vibrant academic cities, a place where historical depth, cultural continuity, and contemporary scientific exchange meet. Braga offers an inspiring European context for renewing the international and intercultural dialogue that has always characterized ICCSA, while giving new visibility to the role of computational science in addressing the complex challenges of cities, territories, technologies, and societies.

ICCSA 2026 marked another fruitful and thought-provoking academic event in the International Conferences on Computational Science and Its Applications (ICCSA) conference series, previously held in Istanbul, Türkiye (2025), Hanoi, Vietnam (2024), Athens, Greece (2023), Málaga, Spain (2022), Cagliari, Italy (hybrid with few participants in presence in 2021 and completely online in 2020), whilst earlier editions took place in Saint Petersburg, Russia (2019), Melbourne, Australia (2018), Trieste, Italy (2017), Beijing, China (2016), Banff, Canada (2015), Guimarães, Portugal (2014), Ho Chi Minh City, Vietnam (2013), Salvador, Brazil (2012), Santander, Spain (2011), Fukuoka, Japan (2010), Suwon, South Korea (2009), Perugia, Italy (2008), Kuala Lumpur, Malaysia (2007), Glasgow, UK (2006), Singapore (2005), Assisi, Italy (2004), Montreal, Canada (2003), and (as ICCS) Amsterdam, the Netherlands (2002) and San Francisco, USA (2001).

Computational Science constitutes the main pillar of most present research, industrial and commercial applications, and plays a unique role in exploiting innovative ICT technologies, and the ICCSA conference series has, accordingly, provided ample opportunities to researchers and industry practitioners to discuss new ideas, to share complex problems and their solutions, and to shape new trends in Computational Science. As the conference mirrors society from a scientific point of view, this year's undoubtedly

dominant theme was large language models, machine learning, and Artificial Intelligence (AI), and their applications across the most diverse technological, economic, and industrial fields, amongst others.

The ICCSA 2026 conference was structured into six general tracks covering computational science and its applications: Computational Methods, Algorithms and Scientific Applications; High-Performance Computing and Networks; Geometric Modelling, Graphics and Visualisation; Advanced and Emerging Applications; Information Systems and Technologies; and Urban and Regional Planning. In addition, the conference included 57 workshops, focusing on topical issues of utmost importance to science, technology, and society: from new computational approaches for earth science, to mathematical methods for image processing, new statistical and optimisation methods, several Artificial Intelligence approaches, sustainability issues, smart cities and related technologies, to name some.

In the Main Conference Proceedings, we accepted 119 full papers, 3 short papers and 2 Ph.D. Showcase papers on 405 submissions to the General Tracks of the Conference (with an acceptance rate of 29.9%). Reviews were single blind, and submissions received 2.6 reviews each, on average. In the Workshops proceedings, we accepted 321 full papers, 23 short papers and 3 Ph.D. Showcase papers from a total of 1184 submissions (Acceptance rate 29.9%). We would like to convey our sincere appreciation to the workshop chairs and co-chairs, and program committee members, for their diligent work, commitment, and dedication.

The success and consistent maintenance of the ICCSA conference series in general, and of ICCSA 2026 in particular, rely upon the support of many people: authors, presenters, participants, keynote speakers, workshop chairs, session chairs, Organizing Committee members, student volunteers, Program Committee members, Advisory Committee members, International Liaison chairs, reviewers, and other individuals in various roles. Thus, we take this opportunity to wholeheartedly thank each and every one.

We also wish to thank the publisher, Springer, for accepting the proceedings, for sponsoring part of the best paper awards, and for their kind assistance and cooperation during the editing process. We would cordially like to invite you to refer to the ICCSA website https://iccsa.org, where you can find the relevant details regarding this academic endeavour and our event.

June 2026

Chiara Garau
Ana Maria A. C. Rocha
Osvaldo Gervasi
Beniamino Murgante

Welcome Message from the Organizers

On behalf of the Local Organizing Committee of ICCSA 2026, it is our great pleasure to welcome you to the 26th International Conference on Computational Science and Its Applications, held from June 30 to July 3, 2026, at the University of Minho in Braga, Portugal.

We are deeply honoured by the trust placed in us by the ICCSA General Chairs in organizing another edition of this successful conference series. We hope that ICCSA 2026 provided an inspiring environment for scientific exchange, collaboration, and discussion among researchers from all over the world.

The conference took place at the University of Minho, one of the most prestigious higher education institutions in Portugal. Founded in 1973, the University of Minho has campuses in Braga and Guimarães and is widely recognized for its academic excellence, research activity, and strong international outlook.

Braga, one of the oldest cities in Portugal, is renowned for its historical heritage, cultural vitality, and welcoming atmosphere. Located in the heart of the Minho region, the city combines tradition with modernity, offering visitors renowned gastronomy, beautiful landscapes, and the unique character of *Vinho Verde* wine.

The scientific program of ICCSA 2026 included plenary lectures delivered by distinguished scientists, together with a broad range of workshops covering several areas of Computational Science and its Applications. We believe that the conference provided excellent opportunities to share ideas, discuss emerging challenges, and foster new collaborations.

In addition to the scientific sessions, participants had the opportunity to enjoy a rich social program. Planned activities included a Welcome Reception at Largo do Paço, the Rectorate of the University of Minho, located in the former Archbishop's Palace - one of the most emblematic historical buildings in the city center; a *Vinho Verde* wine tasting experience at Quinta d'Amares, where you can discover the art of winemaking; and a guided visit to the coastal city of Viana do Castelo, renowned for its maritime heritage, traditional architecture, and beautiful Atlantic scenery. The conference gala dinner took place at Colunata Eventos, located within the Bom Jesus do Monte sanctuary, a UNESCO World Heritage Site, where ICCSA participants enjoyed delicious recipes and a wonderful view over the city of Braga.

The success of ICCSA 2026 would not have been possible without the dedication and enthusiasm of many collaborators and volunteers, whose hard work and commitment greatly contributed to the organization of the conference. On behalf of the Local Organizing Committee, it was our honour to warmly welcome you to the beautiful city of Braga and we hope you had an enjoyable and productive conference.

Sincerely yours,

On behalf of the Local Organizing Committee

Ana Maria A. C. Rocha

Organization

Honorary General Chairs

Bernady O. Apduhan	Kyushu Sangyo University, Japan
Kenneth C. J. Tan	Sardina Systems, UK

General Chairs

Ana Maria A. C. Rocha	University of Minho, Portugal
David Taniar	Monash University, Australia
Osvaldo Gervasi	University of Perugia, Italy

Program Committee Chairs

Beniamino Murgante	University of Basilicata, Italy
Chiara Garau	University of Cagliari, Italy

International Advisory Committee

Jemal Abawajy	Deakin University, Australia
Dharma P. Agarwal	University of Cincinnati, USA
Jagannath Aryal	University of Melbourne, Australia
Rajkumar Buyya	University of Melbourne, Australia
Claudia Bauzer Medeiros	University of Campinas, Brazil
Carl K. Chang	Iowa State University, USA
Elizabeth Chang	University of the Sunshine Coast, Australia
Manfred M. Fisher	Vienna University of Economics and Business, Austria
Pierre Frankhauser	University of Franche-Comté/CNRS, France
Marina L. Gavrilova	University of Calgary, Canada
Sumi Helal	University of Florida, USA & Lancaster University, UK
Bin Jiang	University of Gävle, Sweden
Yee Leung	Chinese University of Hong Kong, China
Biswajeet Pradhan	University of Technology Sydney, Australia
My Thai	University of Florida, USA

International Liaison Chairs

Giuseppe Borruso	University of Trieste, Italy
Elise De Donker	Western Michigan University, USA
Maria Noelia Faginas Lago	University of Perugia, Italy
Maria Irene Falcão	University of Minho, Portugal
Robert C. H. Hsu	Chung Hua University, Taiwan
Yeliz Karaca	UMass Chan Medical School, USA
Tae-Hoon Kim	Zhejiang University of Science and Technology, China
Vladimir Korkhov	Saint Petersburg University, Russia
Takashi Naka	Kyushu Sangyo University, Japan
Rafael D. C. Santos	National Institute for Space Research, Brazil
Maribel Yasmina Santos	University of Minho, Portugal
Anastasia Stratigea	National Technical University of Athens, Greece
Nguyen Canh Thai	Thuyloi University, Vietnam
Kh Md Nahiduzzaman	Mohammed VI Polytechnic University, Rabat, Morocco

Workshop and Session Organizing Chairs

Beniamino Murgante	University of Basilicata, Italy
Chiara Garau	University of Cagliari, Italy

Award Chair

Wenny Rahayu	La Trobe University, Australia

Publicity Committee Chairs

Elmer Dadios	De La Salle University, Philippines
Nataliia Kulabukhova	Saint Petersburg University, Russia
Daisuke Takahashi	Tsukuba University, Japan
Doan Quang Tu	Academy of Military Science and Technology, Vietnam
Shangwang Wang	Beijing University of Posts and Telecommunications, China

Local Organizing Committee Chairs

Maria Irene Falcão	University of Minho, Portugal
Ana Cristina Braga	University of Minho, Portugal
Rui José	University of Minho, Portugal
José Manuel Machado	University of Minho, Portugal
Jorge G. Rocha	University of Minho, Portugal

Technology Chair

Damiano Perri	University of Perugia, Italy

Program Committee

Vera Afreixo	University of Aveiro, Portugal
Vladimir Alarcon	Northern Gulf Institute, USA
Filipe Alvelos	University of Minho, Portugal
Debora Anelli	Polytechnic University of Bari, Italy
Muhammad Waqas Arshad	University of Bologna, Italy
Vittorio Astarita	University of Calabria, Italy
Hartmut Asche	Hasso-Plattner-Institut für Digital Engineering Ggmbh, Germany
Nizamettin Aydın	İstanbul Technical University, Türkiye
Ginevra Balletto	University of Cagliari, Italy
Nadia Balucani	University of Perugia, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
David Berti	ART SpA, Italy
Sandro Bimonte	CEMAGREF, TSCF, France
Ana Cristina Braga	University of Minho, Portugal
Tiziana Campisi	University of Enna KORE, Italy
Alessandra Capolupo	Polytechnic of Bari, Italy
Leocadio G. Casado	University of Almeria, Spain
Ilaria Catapano	National Research Council, IREA, Italy
Maria Cerreta	University of Naples Federico II, Italy
Rachel Chien-Sing Lee	Sunway University, Malaysia
Birol Ciloglugil	Ege University, Turkey
Florbela Maria da Cruz Domingues Correia	Polytechnic Institute of Viana do Castelo, Portugal
Alessandro Costantini	INFN, Italy
Roberto De Lotto	University of Pavia, Italy
Luiza De Macedo Mourelle	State University of Rio De Janeiro, Brazil

Marcelo De Paiva Guimaraes	Federal University of São Paulo, Brazil
Marco De Santis	University of Salerno, Italy
Frank Devai	London South Bank University, UK
Joana Matos Dias	University of Coimbra, Portugal
Aziz Dursun	Virginia Tech University, USA
Laila El Ghandour	Heriot Watt University, UK
Maria Irene Falcão	University of Minho, Portugal
Florbela P. Fernandes	Polytechnic Institute of Bragança, Portugal
Paula Odete Fernandes	Polytechnic Institute of Bragança, Portugal
Adelaide de Fátima Baptista Valente Freitas	University of Aveiro, Portugal
Andreas Fricke	University of Potsdam, Germany
Raffaele Garrisi	Centro Operativo per la Sicurezza Cibernetica, Italy
Ivan Gerace	University of Perugia, Italy
Maria Giaoutzi:	National Technical University of Athens, Greece
Salvatore Giuffrida	University of Catania, Italy
Teresa Guarda	Universidad Estatal Península de Santa Elena, Ecuador
Gabriel Guerrero Contreras	University of Cádiz, Spain
Malgorzata Hanzl	Technical University of Lodz, Poland
Maulana Adhinugraha Kiki	Telkom University, Indonesia
JungYoon Kim	Gachon University, South Korea
Clement Ho Cheung Leung	Chinese University of Hong Kong, China
Andrea Lombardi	University of Perugia, Italy
Marcos Mandado Alonso	University of Vigo, Spain
Ernesto Marcheggiani	Katholieke Universiteit Leuven, Belgium
Barbara Martelli	INFN, Italy
Antonino Marvuglia	Luxembourg Institute of Science and Technology, Luxembourg
Michele Mastroianni	University of Salerno, Italy
Hideo Matsufuru	High Energy Accelerator Research Organization, Japan
Fernando Miranda	Universidade do Minho, Portugal
Giuseppe Modica	University of Reggio Calabria, Italy
Majaz Moonis	UMass Chan Medical School, USA
Leila Moradi	University of Salerno, Italy
Nadia Nedjah	State University of Rio de Janeiro, Brazil
Paolo Nesi	University of Florence, Italy
Suzan Obaiys	University of Malaya, Malaysia
Marcin Paprzycki	Polish Academy of Sciences, Poland
Eric Pardede	La Trobe University, Australia

Ana Isabel Pereira	Polytechnic Institute of Bragança, Portugal
Damiano Perri	University of Perugia, Italy
Massimiliano Petri	University of Pisa, Italy
Telmo Pinto	University of Coimbra, Portugal
Maurizio Pollino	ENEA, Italy
Marcos Quiles	Federal University of São Paulo, Brazil
Nguyen Huu Quynh	Thuyloi University, Vietnam
Albert Rimola	Universitat Autònoma de Barcelona, Spain
Humberto Rocha	University of Coimbra, Portugal
Marzio Rosi	University of Perugia, Italy
Lucia Saganeiti	University of L'Aquila, Italy
Francesco Scorza	University of Basilicata, Italy
Marco Paulo Seabra dos Reis	University of Coimbra, Portugal
Jie Shen	University of Michigan, USA
Bruno Sousa	University of Coimbra, Portugal
Francesco Tajani	Sapienza University of Rome, Italy
Daniela Tavano	University of Calabria, Italy
Ho Sy Tam	Thuyloi University, Vietnam
Deodato Tapete	Italian Space Agency ASI, Italy
Rodrigo Tapia-McClung	Centro de Investigación en Ciencias de Información Geoespacial, Mexico
Tarek Teba	University of Portsmouth, UK
Ana Paula Teixeira	Universidade do Minho, Portugal
Luis Teodoro	University of Oslo, Norway
Yiota Theodora	National Technical University of Athens, Greece
Elena Todella	Polytechnic University of Turin, Italy
Graca Tomaz	Instituto Politécnico da Guarda, Portugal
Anna Tonazzini	National Research Council, Italy
Francesca Torrieri	Polytechnic University of Milan, Italy
Vincenza Torrisi	University of Catania, Italy
Vincenzo Totaro	Polytechnic University of Bari, Italy
Arianna Travaglini	University of Basilicata, Italy
Stefanos Tsigdinos	National Technical University of Athens, Greece
Aime Tsinda	University of Rwanda, Rwanda
Elen Twrdy	University of Ljubljana, Slovenia
Doan Quang Tu	Academy of Military Science and Technology, Vietnam
Le Quang Tuan	Thuyloi University, Vietnam
Toshihiro Uchibayashi	Kyushu University, Japan
Fukuko Yuasa	High Energy Accelerator Research Organization, Japan
Ljiljana Zivkovic	Republic Geodetic Authority, Serbia

General Tracks

1. Computational Methods, Algorithms and Scientific Applications

Computational Biology
Computational Combustion
Computational Chemistry
Computational Fluid Dynamics
Computational Physics
Computational Geometry
Computational Mathematics
Computational Mechanics
Computational Electro-magnetics
Numerical Methods and Algorithms

2. Track: High Performance Computing and Networks

Parallel and Distributed Computing
Cluster Computing
Supercomputing
Cloud Computing
Autonomic Computing
P2P Computing
Mobile Computing
Edge Computing
Workflow Design and Practice
Computer and Network Architecture

3. Track: Geometric Modeling, Graphics and Visualization

Scientific Visualization
Computer Graphics
Geometric Modeling
Pattern Recognition
Image Processing
CAD/CAM
Web3D, Virtual and Augmented Reality

4. Track: Advanced and Emerging Applications

Biochemistry
Bioinformatics
Astrophysics

Biometric Modeling
Environmental, Climate and Weather Modeling
Geology and Geophysics
Nuclear Physics
Financial and Economical Modeling
Computational Journalism

5. Track: Information Systems and Technologies

Information Retrieval
Scientific Databases
Security Engineering
Risk Analysis
Reliability Engineering
Software Engineering
Data Mining
Artificial Intelligence
Machine Learning
Learning Technologies
Web Based Computing
Web 2.0
Blockchain

6. Track: Urban and Regional Planning

Urban and regional growth
Sustainable urban and regional development
Socio-ecological systems
Open Data- Big Data
Cultural Heritage
Smart and Sustainable Cities
Mobility and Intelligent Transport Systems
Geographical Information Systems
Decision Support Systems
Complexity assessment and mapping
Logistics

Workshops

Workshop on Advancements in Applied Machine-Learning and Data Analytics (AAMDA 2026)

Workshop Organizers:

Alessandro Costantini	INFN, Italy
Elisabetta Ronchieri	INFN CNAF, Italy

Artificial Intelligence Tools In Managing Future Communities (AiMFutCOM 2026)

Workshop Organizers:

Marco Vona	University of Basilicata, Italy
Francesco Scorza	University of Basilicata, Italy
Donatello Cardone	University of Basilicata, Italy
Maurizio Pollino	ENEA, Italy
Gregorio DAgostino	ENEA, Italy

Advanced and Innovative Web Apps 2026 (AIWA 2026)

Workshop Organizers:

Damiano Perri	University of Perugia, Italy
Osvaldo Gervasi	University of Perugia, Italy
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy

Advanced Mathematics and Computing Models in Complex Data-Intensive Computational Systems (AMCM 2026)

Workshop Organizers:

Yeliz Karaca	UMass Chan Medical School and Massachusetts Institute of Technology, USA
Dumitru Baleanu	Lebanese American University, Lebanon
Osvaldo Gervasi	University of Perugia, Italy
Yudong Zhang	University of Leicester, UK
Majaz Moonis	UMass Chan Medical School, USA

Advanced Numerical Approaches for Assessment and Design of No-Tension Masonry Structures (ANAMS 2026)

Workshop Organizers:

Antonino Iannuzzo	Universitá degli studi del Sannio, Italy
Natalia Pingaro	Politecnico di Milano, Italy
Andrea Montanino	International Centre for Numerical Methods in Engineering, Spain
Elham Mousavian	University of Edinburgh, UK
Stefano Galassi	University of Florence, Italy

Artificialization of land, Urban Transformation, and Territorial Resilience in a Changing Climate (AUTTREC 2026)

Workshop Organizers:

Celestina Fazia	Università di Enna Kore, Italy
Kh Md Nahiduzzaman	Mohammed VI Polytechnic University, Morocco
Hassan Radoine	Mohammed VI Polytechnic University, Morocco
Cezar Morar	University of Oradea, Romania
Antonio García Martínez	Universidad de Sevilla, Spain
Clara Stella Vicari Aversa	Università degli Studi Mediterranea di Reggio Calabria, Italy

Blockchain and Distributed Ledgers: Technologies and Applications (BDLTA 2026)

Workshop Organizers:

Vladimir Korkhov	Saint Petersburg State University, Russia
Elena Stankova	Saint Petersburg State University, Russia
Nataliia Kulabukhova	AutoTech, Russia

Bio and Neuro inspired Computing and Applications (BIONCA 2026)

Workshop Organizers:

Nadia Nedjah	State University of Rio de Janeiro, Brazil
Luiza de Macedo Mourelle	State University de Rio de Janeiro, Brazil

Computational and Applied Mathematics (CAM 2026)

Workshop Organizers:

Maria Irene Falcao	University of Minho, Portugal
Fernando Miranda	Universidade do Minho, Portugal

Computational and Applied Statistics for Data and AI (CAS 2026)

Workshop Organizers:

Ana Cristina Braga	University of Minho, Portugal
Ines Sena	University of Minho, Portugal
Joao Mendes	Polytechnic Institute of Bragança, Portugal

Circular Economy and Territorial Competitiveness (CETC 2026)

Workshop Organizers:

Gerardo Carpentieri	University of Naples Federico II, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Zhadyra Konurbaeva	D. Serikbayev East Kazakhstan Technical University, Kazakhstan
Saule Kumargazhanova	D.Serikbayev East Kazakhstan Technical University, Kazakhstan

Circular Ecological Heritage Communities: Regeneration and Evaluation Strategies for Historic Villages, Landscapes and Rural Territories (CHEC 2026)

Workshop Organizers:

Mariarosaria Angrisano	Pegaso University, Italy
Antonia Gravagnuolo	CNR Institute of Heritage Science, Italy
Francesca Buglione	CNR ISPC, Italy
Piera Della Morte	Pegaso Telematic University, Italy

Cyber Intelligence and Applications (CIA 2026)

Workshop Organizers:

Gianni DAngelo	University of Salerno, Italy
Francesco Palmieri	University of Salerno, Italy
Massimo Ficco	University of Salerno, Italy

Computational Methods, Statistics and Industrial Mathematics (CMSIM 2026)

Workshop Organizers:

Maria Filomena Teodoro	CINAV, Escola Naval, Instituto Universitário Militar, Portugal
Paula Simões	Academia Militar, Portugal
Teresa A. Oliveira	University of Lisbon and Universidade Aberta, Portugal

Computational Optimization and Applications (COA 2026)

Workshop Organizers:

Ana Maria A. C. Rocha	University of Minho, Portugal
Humberto Rocha	University of Coimbra, Portugal

Computational Methods in Advanced Networking Technologies (CoMANT 2026)

Workshop Organizers:

Eslam Farsimadan	University of Salerno, Italy
Leila Moradi	University of Salerno, Italy
Francesco Palmieri	University of Salerno, Italy

Computational Astrochemistry 2026 (CompAstro 2026)

Workshop Organizers:

Marzio Rosi	University of Perugia, Italy
Daniela Ascenzi	Università degli Studi di Trento, Italy
Nadia Balucani	University of Perugia, Italy
Stefano Falcinelli	University of Perugia, Italy

Computational Methods for Porous Materials (CompPor 2026)

Workshop Organizers:

Vadim Lisitsa	Institute of Petroleum Geology and Geophysics SB RAS, Russia
Evgeniy Romenski	Sobolev Institute of Mathematics, Russia

Workshop on Computational Science and HPC (CSHPC 2026)

Workshop Organizers:

Elise de Doncker	Western Michigan University, USA
Hideo Matsufuru	High Energy Accelerator Research Organization, Japan
Robert Makin	Western Michigan University, USA

Cities, Technologies and Planning (CTP 2026)

Workshop Organizers:

Giuseppe Borruso	University of Trieste, Italy
Beniamino Murgante	University of Basilicata, Italy
Malgorzata Hanzl	Lodz University of Technology, Poland
Anastasia Stratigea	National Technical University of Athens, Greece
Ljiljana Zivkovic	Republic Geodetic Authority, Serbia
Ginevra Balletto	University of Cagliari, Italy

Evaluating Inner Areas Potentials (EIAP 2026)

Workshop Organizers:

Diana Rolando	Politecnico di Torino, Italy
Manuela Rebaudengo	Politecnico di Torino, Italy
Alice Barreca	Politecnico di Torino, Italy
Giorgia Malavasi	Politecnico di Torino, Italy

sustainablE mobiLity and Last mIle lOgisTic in Urban Context (ELLIOT 2026)

Workshop Organizers:

Tiziana Campisi	University of Enna Kore, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
Florin Rusca	National University of Science and Technology Politehnica Bucharest, Romania
Grigorios Fountas	Aristotle University of Thessaloniki, Greece
Aleksandra Deluka-Tibljas	University of Rijeka, Croatia

Econometrics and Multidimensional Evaluation in Urban Environment (EMEUE 2026)

Workshop Organizers:

Maria Cerreta	University of Naples Federico II, Italy
Giuseppe Ciciriello	University of Naples Federico II, Italy
Simona Panaro	University of Sussex Business School, UK
Piero Zizzania	Università of Naples Federico II, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
F rancesco Tajani	Sapienza University of Rome, Italy

Ecosystem Services in Spacial Planning for Climate Neutral Urban and Rural Areas (ESSP 2026)

Workshop Organizers:

Sabrina Lai	University of Cagliari. Italy
Carmela Gargiulo	Università Federico II di Napoli, Italy
Beniamino Murgante	University of Basilicata, Italy
Francesco Scorza	University of Basilicata, Italy
Corrado Zoppi	University of Cagliari, Italy
Floriana Zucaro	University of Naples Federico II, Italy

The 16th International Workshop on Future Information System Technologies and Applications (FiSTA 2026)

Workshop Organizers:

Bernady O. Apduhan	Kyushu Sangyo University, Japan
Rafael Santos	National Institute for Space Research (INPE), Brazil

Geographical Analysis, Urban Modeling, Spatial Statistics (Geog-And-Mod 2026)

Workshop Organizers:

Giuseppe Borruso	University of Trieste, Italy
Beniamino Murgante	University of Basilicata, Italy
Hartmut Asche	Hasso Plattner Institute for Digital Engineering, Germany
Rodrigo Tapia McClung	Centro de Investigación en Ciencias de Información Geoespacial, Mexico
Andreas Fricke	Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany

Generative and Neuro-Symbolic Artificial Intelligence for Real-World Systems (GNS-AI 2026)

Workshop Organizers:

Hugo Peixoto	University of Minho, Portugal
Manuel Rodrigues	University of Minho, Portugal
Regina Sousa	University of Minho, Portugal

Geomatics for Resource Monitoring and Management (GRMM 2026)

Workshop Organizers:

Alberico Sonnessa	Politecnico di Bari, Italy
Eufemia Tarantino	Politecnico di Bari, Italy
Marco Locurcio	Polytechnic University of Bari, Italy

International Workshop on Information and Knowledge in the Internet of Things (IKIT 2026)

Workshop Organizers:

Teresa Guarda	Universidad Estatal Península de Santa Elena, Ecuador

Intelligent Sustainable Development: from Models to Real-World Impact (ISD 2026)

Workshop Organizers:

Diego Brandão	Federal Center for Technological Education of Rio de Janeiro, Brazil
Angela Ferreira	Polytechnic University of Bragança, Portugal
Eduardo Bezerra	Federal Center for Technological Education of Rio de Janeiro, Brazil
Raphael Guerra	Universidade Federal Fluminense, Brazil
Douglas Cardoso	University of Coimbra, Portugal
Rui Pedro Lopes	Polytechnic Institute of Bragança, Portugal

Multidimensional Evolutionary Evaluations for Transformative Approaches (MEETA 2026)

Workshop Organizers:

Maria Cerreta	University of Naples Federico II, Italy
Giuliano Poli	Università Federico II di Napoli, Italy
Gaia Daldanise	CNR Institute for Research on Innovation and Services for Development, Italy
Ludovica La Rocca	University of Federico II of Naples, Italy
Maria Somma	University of Naples Federico II, Italy
Sabrina Sacco	University of Naples Federico II, Italy

Building Multi-dimensional Models for Assessing Complex Environmental Systems (MES 2026)

Workshop Organizers:

Vanessa Assumma	University of Bologna, Italy
Caterina Caprioli	Politecnico di Torino, Italy
Giulia Datola	Politecnico di Milano, Italy
Federico DellAnna	Politecnico di Torino, Italy
Marta DellOvo	Politecnico di Milano, Italy
Marco Rossitti	Politecnico di Milano, Italy

Computational Mechanics, Intelligence, and Data (MIND 2026)

Workshop Organizers:

Lelio Campanile	Università degli Studi della Campania Luigi Vanvitelli, Italy
Renato Zona	Università della Campania Vanvitelli, Italy

Modelling Liveable Cities: Techniques, Methods, Challenges, and Perspectives Behind the X-Minute City (MLC 2026)

Workshop Organizers:

Federico Mara	University of Pisa, Italy
Valerio Cutini	University of Pisa, Italy

Mathematical Methods for Image Processing and Understanding 2026 (MMIPU 2026)

Workshop Organizers:

Ivan Gerace	Università degli Studi di Perugia, Italy
Gianluca Vinti	Università degli Studi di Perugia, Italy
Arianna Travaglini	Lucian Blaga University of Sibiu, Romania

MOdels and Indicators for Assessing and Measuring the Urban Settlement deVElopment in the View of NET ZERO by 2050 (MOVEto0 2026)

Workshop Organizers:

Lorena Fiorini	University of L'Aquila, Italy
Lucia Saganeiti	Institute of Methodologies for Environmental Analysis (IMAA) CNR, Italy
Angela Pilogallo	Institute of Methodologies for Environmental Analysis (IMAA), Italy
Francesco Zullo	University of L'Aquila, Italy
Alessandro Marucci	University of L'Aquila, Italy

Numerical Aspects in Machine Learning for Dynamical Systems (NAMLYS 2026)

Workshop Organizers:

Pasquale De Luca	Parthenope University of Naples, Italy
Livia Marcellino	University of Naples Parthenope, Italy
Valeria Mele	University of Naples Federico II, Italy

Numerical Methods and Machine Learning in Science and Engineering (NMLSE 2026)

Workshop Organizers:

Cecília Coelho	Helmut Schmidt University, Germany
M. Fernanda P. Costa	University of Minho, Portugal
Luís Ferrás	University of Minho, Portugal
Eliana Costa e Silva	Polytechnic of Porto, Portugal

6th Workshop on Privacy in the Cloud/Edge/IoT World 2026 (PCEIoT 2026)

Workshop Organizers:

Lelio Campanile	Università degli Studi della Campania Luigi Vanvitelli, Italy

Mauro Iacono	Università degli Studi della Campania Luigi Vanvitelli, Italy
Michele Mastroianni	University of Foggia, Italy

Preserving Our Past: Spatial and Remote Sensing Technologies for Cultural Heritage in a Changing Climate (POP 2026)

Workshop Organizers:

Maria Danese	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy
Rosa Lasaponara	Institute of Methodologies for Environmental Analysis (CNR IMAA), Italy
Nicola Masini	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy

Scientific Computing Infrastructure (SCI 2026)

Workshop Organizers:

Vladimir Korkhov	Saint Petersburg State University, Russia
Elena Stankova	Saint Petersburg State University, Russia

Socio-Economic and Environmental Models for Land Use Management (SEMLUM 2026)

Workshop Organizers:

Debora Anelli	Sapienza University of Rome, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
Francesco Tajani	Sapienza University of Rome, Italy
Felicia Di Liddo	Polytechnic University of Bari, Italy

Ports and Logistics of the Future - Smartness and Sustainability (SmartPorts 2026)

Workshop Organizers:

Andrea Gallo	Università degli Studi di Trieste, Italy
Giuseppe Borruso	University of Trieste, Italy

Gianfranco Fancello	University of Cagliari, Italy
Veronica Camerada	Università degli studi di Sassari, Italy
Clara Di Fazio	University of Trieste, Italy

Sustainable Evolution of Long-Distance frEight and paSsenger Transport (SOLIDEST 2026)

Workshop Organizers:

Antonio Russo	University of Enna Kore, Italy
Giuseppe Musolino	Università Mediterranea di Reggio Calabria, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
Oana Dinu	National University of Science and Technology Politehnica Bucharest, Romania
Sanja Šurdonja	University of Rijeka, Croatia
Elen Twrdy	University of Ljubljana, Slovenia

Sustainability Performance Assessment: Models, Approaches, and Applications Toward Interdisciplinary and Integrated Solutions. (SPA 2026)

Workshop Organizers:

Francesco Scorza	University of Basilicata, Italy
Sabrina Lai	University of Cagliari, Italy
Jolanta Dvarioniene	Kaunas University of Technology, Lithuania
Michele Campagna	University of Cagliari, Italy

Specifics of Smart Cities Development in Europe (SPEED 2026)

Workshop Organizers:

Chiara Garau	University of Cagliari, Italy
Katarína Vitálišová	Matej Bel University, Slovak Republic
Marco Fanfani	University of Florence, Italy
Anna Vaňová	Matej Bel University, Slovak Republic
Kamila Borsekova	Matej Bel University, Slovak Republic
Paola Zamperlin	University of Pisa, Italy

Smart, Safe, and Health Cities (SSHC 2026)

Workshop Organizers:

Chiara Garau	University of Cagliari, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Martina Carra	University of Brescia, Italy
Vincenza Torrisi	University of Catania, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy

Smart and Sustainable Island Communities (SSIC 2026)

Workshop Organizers:

Chiara Garau	University of Cagliari, Italy
Anastasia Stratigea	National Technical University of Athens, Greece
Yiota Theodora	National Technical University of Athens, Greece
Giovanna Concu	University of Cagliari, Italy
Tiziana Campisi	University of Enna Kore, Italy
Tanja Congiu	University of Sassari, Italy

Sustainable Tourism Economy, Silk Road, Reuse and Enhancement of Cultural Heritage (STESilkRoad 2026)

Workshop Organizers:

Celestina Fazia	Università di Enna Kore, Italy
Alexandru Ilies	University of Oradea, Romania
Dorina Camelia Ilies	University of Oradea, Romania
Cristina Natoli	Ministero dell'Ambiente e della Sicurezza Energetica, Italy
Bahodirhon Safarov	Samarkand Branch of Tashkent State University of Economics, Uzbekistan
Clara Stella Vicari Aversa	Università degli Studi Mediterranea di Reggio Calabria, Italy

Theoretical and Computational Chemistry and its Applications (TCCM 2026)

Workshop Organizers:

Noelia Faginas Lago	Università di Perugia, Italy
Andrea Lombardi	Università di Perugia, Italy
Marcos Mandado Alonso	University of Vigo, Spain

Transport Infrastructures for Smart Cities (TISC 2026)

Workshop Organizers:

Francesca Maltinti	University of Cagliari, Italy
Mauro Coni	University of Cagliari, Italy
Chiara Garau	University of Cagliari, Italy
Benedetto Barabino	University of Brescia, Italy
Nicoletta Rassu	Università di Cagliari, Italy
James Rombi	University of Cagliari, Italy

Transforming Urban Analytics: The Impact of Crowdsourced Mapping and Advanced AI Techniques on Future Cities (Tr-UrbAna 2026)

Workshop Organizers:

Ayse Giz Gulnerman Gengec	Ankara Hacı Bayram Veli University, Turkey
Müslüm Hacar	Yildiz Technical University, Turkey

From Structural to TRAnsformative-Change of City Environment: Challenges and Solutions and Perspectives (TRACE 2026)

Workshop Organizers:

Maria Rosaria Guarini	Sapienza University of Rome, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
Francesco Sica	Sapienza University of Rome, Italy
Alejandro Segura	Universidad Politécnica de Madrid, Spain
Yilsy María Nuñez Guerrero	Universidad Politécnica de Madrid, Spain

Urban Artificial Intelligence for Modelling, Analysis and Planning (Urban plAnnIng 2026)

Workshop Organizers:

Chiara Garau	University of Cagliari, Italy
Romano Fistola	University of Naples Federico II, Italy
Beniamino Murgante	University of Basilicata, Italy
Dessislava Petrova Antonova	Sofia University, GATE Institute, Bulgaria
Gianni Pantaleo	University of Florence, Italy
Rosa Anna La Rocca	University of Naples Federico II, Italy

Urban Space Accessibility and Mobilities (USAM 2026)

Workshop Organizers:

Chiara Garau	University of Cagliari, Italy
Alessandro Plaisant	University of Sassari, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Mauro D. Apuzzo	University of Cassino and Southern Lazio, Italy
Francesco Pinna	University of Cagliari, Italy
Silvio Nocera	IUAV University of Venice, Italy

Virtual Reality, Augmented Reality and Applications (VRA 2026)

Workshop Organizers:

Damiano Perri	University of Perugia, Italy
Osvaldo Gervasi	University of Perugia, Italy
Paolo Nesi	University of Florence, Italy

Workshop on Advanced and Computational Methods for Earth Science Applications (WACM4ES 2026)

Workshop Organizers:

Luca Piroddi	University of Cagliari, Italy
Patrizia Capizzi	University of Palermo, Italy
Marilena Cozzolino	Università del Molise, Italy
Giuseppina Vacca	University of Cagliari, Italy
Luigi Capozzoli	Hydrogeosite Laboratory of CNR-IMAA, Italy
Nikos Papadopoulos	Foundation for Research and Technology Hellas, Greece

Sponsoring Organizations

ICCSA 2026 would not have been possible without the tremendous support of many organizations and institutions, for which all organizers and participants of ICCSA 2026 express their sincere gratitude:

Universidade do Minho
Escola de Engenharia

University of Minho, Portugal (https://www.uminho.pt/)

University of Perugia, Italy (https://www.unipg.it)

University of Basilicata, Italy (http://www.unibas.it)

University of Cagliari, Italy (https://unica.it/)

Monash University, Australia (https://www.monash.edu/)

Kyushu Sangyo University, Japan (https://www.kyusan-u.ac.jp/)

Venue

ICCSA 2026 took place at:

the University of Minho campus:
Campus de Gualtar
4710-057 Braga
Portugal

Plenary Lectures

Digital Twins for Cities – The Rise of the Socio-Technical System

Claudia van der Laag

Professor of Urban Analytics, Department of Built Environment, Oslo Metropolitan University, Oslo

Abstract. Cities are entering a new era in which digital twins are no longer simply technical replicas of urban systems, but evolving socio-technical infrastructures that shape how cities are planned, governed, and experienced. Across the world, urban digital twins are being adopted to address some of the most pressing challenges of our time—climate adaptation, urban resilience, mobility transitions, decarbonisation, infrastructure management, and inclusive governance. Yet as these technologies become increasingly sophisticated, a fundamental question emerges: can digital twins truly support more democratic and human-centred cities, or will they reinforce existing structures of power and expertise? In this keynote, Claudia van der Laag, explores the transformation of urban digital twins from data-driven modelling tools into collaborative platforms for urban decision-making and societal change. It argues that the future of digital twins lies not only in advances in artificial intelligence, real-time data integration, simulation, and predictive analytics, but equally in the ability to connect technological innovation with governance, participation, and human agency. Drawing on emerging international research and practical case studies, the keynote examines the complex ecosystem surrounding urban digital twins, bringing together planners, policy-makers, researchers, technologists, and communities. It highlights how digital twins increasingly function as meeting points between multiple forms of knowledge—technical, political, social, and experiential—and why interdisciplinary collaboration is essential for their success. The talk will critically reflect on the tensions between technological ambition and civic inclusion, exploring themes such as transparency, interoperability, participation, and trust. It will also discuss how digital twins can become powerful environments for scenario exploration, urban experimentation, and collective learning, enabling cities to better navigate uncertainty and complexity in an age of rapid urban transformation. Ultimately, this keynote presents a vision of digital twins not as neutral technological systems, but as living socio-technical infrastructures capable of reshaping urban governance itself. The challenge ahead is therefore not only how accurately cities can be modelled, but how inclusively they can be imagined, negotiated, and governed through these emerging digital environments.

Short Bio – Claudia van der Laag (also known as Claudia Yamu) is Professor of Urban Analytics at Oslo Metropolitan University. Her work advances urban analytics and spatial modelling through an integrated multi-method approach, informing sustainable urban planning and policy. She is widely recognised for bridging theory and practice and received the prestigious Michael Breheny Prize for her contributions to multi-scale

modelling. Van der Laag is co-author of the widely recognised textbook ‘Introduction to Space Syntax in Urban Studies’. She also serves on the editorial advisory board of Springer’s The Urban Book Series and the journal Environment and Planning B: Urban Analytics and City Science.

Prescriptive AI in Practice: Bridging Optimisation, Generative AI and Agentic Decision-Making

Bernardo Almada-Lobo

Universidade do Porto, Faculdade de Engenharia, INESC TEC, Portugal

Abstract. Prescriptive AI, viewed through the lens of mathematical optimisation and operations research, addresses complex business decision-making problems involving limited resources, multiple decision options and competing objectives. This talk considers the full decision pipeline: problem framing, model formulation, computational solution, implementation, monitoring and organisational adoption. The practical success of prescriptive AI depends less on algorithmic sophistication alone than on the quality of the translation between real-world decision contexts and formal optimisation models. In practice, this translation remains a major bottleneck. Objectives are often ambiguous, constraints are incomplete or tacit, data may be fragmented, and the operational consequences of a recommendation may be poorly understood. As a result, optimisation projects can fail even when the underlying mathematical model is technically sound. The talk will address the persistent gap between optimisation theory and deployment practice. While optimisation has generated substantial value in domains such as supply chain planning, logistics, production, revenue management, workforce planning and inventory control, many initiatives still face a significant risk of under-delivery. Common failure modes include incorrect problem framing, weak data readiness, brittle assumptions, infeasible or non-actionable recommendations, limited explainability, insufficient ownership by business users and inadequate post-deployment monitoring. Crossing this gap requires moving from one-off model building to an operational lifecycle in which models are stress-tested, embedded in workflows, monitored for drift and revised as the business environment changes. Recent developments in generative and agentic AI create new opportunities to reduce these barriers. Generative AI can support the elicitation of decision-maker preferences, the structuring of unstructured operational knowledge, the drafting of candidate formulations and the explanation of model outputs in business language. Agentic AI can extend this role by orchestrating multi-step workflows that involve model formulation, solver-code generation, execution, debugging, infeasibility diagnosis, sensitivity analysis and human review. However, these capabilities should not be interpreted as replacing optimisation expertise. Rather, they create a new interface between human judgement, analytical modelling and computational decision support. This talk will discuss a cross-fertilisation agenda between optimisation, generative AI and agentic AI, with a focus on how these technologies should be calibrated to the decision at hand.

Short Bio – Full Professor of Industrial Engineering and Management at the Faculty of Engineering of University of Porto (FEUP). Co-founder and Partner at LTPlabs, a Business AI company. Member of the Board of Trustees of the Belmiro de Azevedo

Foundation. Member of the Academy of Engineering. Former Board Member of INESC TEC and Researcher at the MIT Sloan School of Management. Holds a BEng, PhD and Habilitation in Industrial Engineering and Management from FEUP. Completed the Advanced Management Programme at INSEAD, and the Mergers and Acquisitions at London Business School. Certified Analytics Professional (CAP) by INFORMS. His area of activity is Business Analytics and Artificial Intelligence. He researches, develops and implements analytical models and methods to support decision-making, solving management problems across various domains—manufacturing, consumer goods, retail and transport—with a particular focus on Operations Management and the Operations–Marketing interface. He has coordinated and participated in more than 300 development, consultancy and research projects funded by private organisations, science agencies from several countries, the European Commission and regional bodies. He has published more than 100 scientific articles and is co-author of the book Analytics Sandwich: Bringing People and Artificial Intelligence Together to Unlock Business Value. He has supervised more than 20 doctoral theses.

Net-Zero Communities Through Life Cycle Thinking: Canadian Lessons and Global Implications

Kasun Hewage

Professor of Engineering, School of Engineering (Okanagan Campus), The University of British Columbia (UBC), Canada, FortisBC Smart Energy Chair, Associate Director, Clean Energy Research Centre (CERC)

Abstract. This keynote presentation discusses how life cycle thinking supports the transformation of communities toward sustainable, resilient, and net-zero futures. Drawing on Canadian research, practice, and policy experience, the presentation highlights how integrated life cycle approaches can move beyond conventional carbon accounting to inform practical decision-making across buildings, infrastructure, energy systems, waste management, water systems, and industrial processes [1-8]. The keynote features research conducted at the Life Cycle Management Laboratory (LCML) at the University of British Columbia, Canada, an internationally recognized research hub led by Prof. Kasun Hewage. LCML brings together engineers, scientists, policy researchers, postdoctoral fellows, visiting scholars, and graduate researchers to translate data-intensive sustainability research into actionable climate solutions. The laboratory's work is organized around six interconnected research themes: energy systems, urban development, water systems, waste management, construction and asset management, and industrial processes and products. Together, these themes provide an integrated platform for advancing life cycle–based pathways to net-zero communities. Supported by strong partnerships with government, industry, and community stakeholders, LCML has contributed to a substantial body of applied sustainability research, including more than 400 journal articles, 100 conference papers, and numerous technical reports that have informed real-world planning and implementation. This keynote presentation highlights selected success stories that demonstrate how life cycle thinking can be translated into practical strategies for reducing environmental impacts while strengthening community resilience. These insights are especially relevant to the interdisciplinary focus of ICCSA 2026, which will be held in collaboration with the University of Minho in Braga, Portugal, from June 30 to July 3, 2026.

References

1. Weerasinghe, O., Hewage, K., Razi, F., Fatmi, M., Sadiq, R.: Applicability of land use models for sustainable urban planning: a review. J. Urban Plann. Dev. ASCE (2025). Accepted 3 Marc 2025

2. Sultana, S.R., et al.: Looking beyond energy efficiency targets: life cycle sustainability of mechanical and water heating equipment in Canadian homes. J. Clean. Prod. **473**, 143469 (2024). https://doi.org/10.1016/j.jclepro.2024.143469
3. Liyanage D.R., Hewage K., Hussain S.H., Razi F., Sadiq R.: Climate adaptation of existing buildings: a critical review on planning energy retrofit strategies for future climate. Renew. Sustain. Energy Rev. **199**, 114476 (2024). https://doi.org/10.1016/j.rser.2024.114476
4. Liyanage D.R., Hewage K., Karunathilake H., Sadiq R.: Feasibility of carbon-capturing in building heating systems: a life cycle thinking-based approach. Int. J. Greenhouse Gas Control **132** (2023). https://doi.org/10.1016/j.ijggc.2023.104056
5. Moudgil, V., Sadiq, R., Bakhtavar, E., Paudel, A., Hewage, K.: A cloud-oriented data-analysis framework to analyze peak demand dynamics in institutional building clusters. Sustain. Cities Soc. **111**, 105553 (2024). https://doi.org/10.1016/j.scs.2024.105553
6. Liyanage D.R., Hewage K., Ghobadi M., Sadiq R.: Thermal resiliency of single-family housing stock under extreme hot and cold conditions. Energy Build. **323**, 114809 (2024). https://doi.org/10.1016/j.enbuild.2024.114809
7. Wijayasekera S.C., et al.: Sustainability of waste-to-hydrogen conversion pathways: a life cycle thinking-based assessment. Energy Convers. Manage. **270** (2022). https://doi.org/10.1016/j.enconman.2022.116218
8. Jayasena A., Hewage K., Siddiqui O., Sadiq R.: Socio-economic and environmental cost-benefit analysis of passive houses: a life cycle perspective. J. Clean. Prod. **373** (2022). https://doi.org/10.1016/j.jclepro.2022.133718

Short Bio – Prof. Kasun Hewage is a Full Professor and FortisBC Smart Energy Chair in the School of Engineering at the University of British Columbia. He also serves as Associate Director of UBC's Clean Energy Research Centre and Director of the Life Cycle Management Laboratory. Prof. Hewage brings extensive multidisciplinary industrial experience in Canada and internationally. Prior to joining UBC in 2008, he worked as a business management cost specialist in the oil and gas sector and as a civil engineer on hydropower and infrastructure development projects. His research focuses on integrating life-cycle thinking into urban development, energy policy, and infrastructure planning. A distinctive feature of his work is its whole-system perspective, extending beyond conventional assessment methods to evaluate products, processes, and built assets from material extraction through end-of-life. Through this approach, Prof. Hewage and his research team have supported governments, infrastructure developers, and communities in advancing low-impact and net-zero development strategies for both existing and planned communities. Prof. Hewage currently leads many collaborative research projects on life-cycle management of built assets, green construction, smart energy systems, and sustainable community planning. His contributions have helped establish important benchmarks, decision-support methods, and implementation pathways for advancing sustainability in the Canadian and global construction sectors. He has authored or co-authored more than 300 peer-reviewed journal publications, in addition to numerous book chapters and conference papers presented at national and international forums. Prof. Hewage is registered as a Professional Engineer in British Columbia, Canada. He is a Fellow of the Canadian Society for Civil Engineering, the Canadian Society of Senior Engineers, and the Engineering Institute of Canada.

Machine Learning Based Methodologies for Grape and Wine Production

Pedro Melo-Pinto

University of Trás-os-Montes and Alto Douro, Engineering Department, Centre for the Research and Technology of Agro-Environmental and Biological Sciences, Portugal

Abstract. Wine production is a major area in Portuguese economy. The sector accounts for approximately 2.6% of Portugal's GDP, and is crucial to the development of inland regions. Wine quality depends on grape quality, which is a result of different factors such as harvesting at the optimal point of maturity, or selecting the grapes according to quality features. The growing importance of precision agriculture, demands novel methods to make the most out of the data generated. Machine learning approaches are particularly suitable for complex processes' modelling, where too many arguments are involved. Many of the agricultural processes such as yield estimation or ripening, fall into this category. However, several challenges may arise: even if models are available for particular cases, general models are difficult to come by, largely due to the high variability involved; with such models, there is a need for huge amounts of data, in a way which will not always be easy to obtain. This is particularly true in the case of grape production and the wine industry, where the number of grape varieties, climate, soil, etc., introduce a degree of variability that is uncommon in other sectors. One way of obtaining more information in such cases has been the use of images, whether RGB, multi-spectral or hyperspectral. The use of images introduces additional challenges, such as data size and lighting conditions, which underscores the need for data-driven modelling methodologies. The use of machine learning (ML) techniques combined with images has been establishing itself as a viable alternative predicting yields or quality parameters in grapes, and assist on harvesting critical decisions. Several ML approaches have been proposed to handle such data characteristics, but selecting a suitable methodology that best address the problem under study and make sure it generalizes well, is, most of the times, a cumbersome task. We will address three wine and grape production problems, yield estimation, grape quality assessment and illness detection, presenting different difficulties. Yield estimation is a major aspect in the wine production industry. The number of berries is one of the most relevant yield aspects, and can be estimated using computer vision, namely image segmentation algorithms. Lighting is one of the key factors in computer vision approaches, and occlusions—which are common in this problem—can be minimized, but require estimation models to achieve reliable results. Illness early detection is, of course, of the utmost importance, with an impact on production and quality. Detecting early symptoms and distinguishing them from other conditions is a difficult task, which makes the use of machine learning algorithms advisable, using automatic visual inspection of the leaves, addressed as a pixel-based classification problem. However, the available datasets are limited and complex to compile, as different diseases typically require very

specific timeframes and locations. The use of generative models to overcome this obstacle, in conjunction with robust classification models, lead to models that are intended to be as general as possible. The evaluation of the grape's maturation profile based on the evolution of their oenological parameters over time is one of the most important aspects for ensuring wine quality, playing a decisive role in the definition of the optimal time for harvesting. The use hyperspectral imaging combined with learning algorithms represents a faster and non-destructive grape ripeness assessment method. We focus in two fundamental aspects to address the resulting data variability: a possible wavelength bands selection (with the purpose of reducing the dimensionality of data without losing predictive power) and the generalization ability of the ML model under such demanding conditions.

Short Bio – Pedro Melo-Pinto graduated from Univ. do Porto in 1984 and received a Ph.D. in Computer Science from the Univ. de Trás-os-Montes e Alto Douro (UTAD) in 1998. He is a Full Professor at UTAD where he served as Pro-Rector, as Director and Vice-Director of R&D units, and as member of UTAD's General Council. He also served as a member of The Portuguese Strategic Infrastructures (for the Digital Areas) Monitoring Committee and usually serves as a member (or co-chair) of the evaluation panel for different Portuguese and international R&D projects and grants competitions. He is currently Associate Editor of Computers and Electronics in Agriculture journal. He was Chair of the EuroFuse 2011 conference and the NATO 2001 Advanced Research Workshop, and served as member of the organization committee for more than 40 international conferences/workshops/special sessions. He is the head of the Image-Based Systems Lab from the Centre for the Research and Technology of Agro-Environmental and Biological Sciences.

His scientific interests include Computer Vision (CV) and Machine Learning (ML) and, in the last 15 years, their applications to the Agro-forestry area, namely to grape oenological parameters assessment. He authored/co-authored more than 100 scientific papers. He participated in many Portuguese and international R&D projects, predominantly in the areas mentioned above.

Contents

General Track 5: Information Systems and Technologies

Gamifying Science Communication: An Experience Report of *Jornada Galáctica* at *Canal Ciência* 3
Cristiane Soares Ramos, Sergio Antônio Andrade de Freitas, Mylena Angélica Silva Farias, Wildemberg Sales da Silva Júnior, Marcus Vinicius Paiva Martins, and Leda Cardoso Sampson Pinto

Mitigating Membership Inference Attacks on Knowledge Graphs 21
Jnaini Abdellah

MobileTestPro: A Real-World Evaluation 38
Michelle Paredes, Danilo Martinez, Mireya Zapata, and Marcelo Rea

ABCD-Based Melanoma Detection in Non-dermoscopic RGB Imaging 53
Maria Oniga, Paul Florin Rus, and Alina Elena Sultana

LEAP: A Leakage-Free Evolutionary Alpha Pipeline for NLP-Driven DJIA Prediction 71
Miguel P. Cunha, Bruno L. Dalmazo, Viviane Leite Dias de Mattos, Richard F. Pinto, Diego Renan Bruno, Eduardo N. Borges, Giancarlo Lucca, Fabian C. Cardoso, and Rafael A. Berri

A High-Throughput Real-Time Object Counting Framework for Industrial Conveyors: Asynchronous Pipelining and Spatiotemporal Consistency Strategy 88
Hieu Trung Nguyen, Thuan Huy Chau, and Vinh Dinh Nguyen

An Integrated Congestion-Aware Framework for Task Allocation and Path Planning in Multi-Agent Pickup and Delivery 100
Dariush Ebrahimi, Meet Sable, Qing Dai, Yash Ahuja, and Michael Ajibola

Bidirectional GRU Neural Network for Simultaneous Human Activity Recognition and Localization Using WiFi CSI Signals 116
Sakorn Mekruksavanich and Anuchit Jitpattanakul

Efficient Satisfiability-Based Approach for Mining Partially-Ordered Sequential Rules 129
Abdelhamid Boudane, Lamia Sadeg, and Mohamed Maiza

Detecting Real Software Faults Through Resource Consumption Anomalies: An Evaluation of the Tricorder Methodology 146
Caio Guimaraes Herrera, Simone R. S. Souza , Ricardo Ribeiro dos Santos , and Paulo Sergio Lopes de Souza

Enhancing the Trustworthiness of Deep Ensemble via Explainability in Medical Imaging 162
Gbègninougbo Aurel Davy Tchokponhoue and Ali Idri

A Laboratory Testbed for Benchmarking Orchestration Logic in Collaborative Environments 177
Leopoldo O. Silva, Diogo L. Rodrigues, Francisco J. Duarte, and Luís S. Dias

Metaheuristic-Optimized Ensemble Learning for Glioma Grading 193
Lucas Silva Santana, Saulo Moraes Villela, Luiz Maurílio da Silva Maciel, Raul Fonseca Neto, and Marcelo Bernardes Vieira

Evaluating an Amazigh Speech Recognition System in a Noisy Environment Using MFDWC Features 209
Fadwa Abakarim and Abdenbi Abenaou

A Heterogeneous Ensemble Meta-Model for Large-Scale Categorization of Portuguese Scientific Records in OasisBR 228
Sergio Antonio Andrade de Freitas, André Corrêa da Silva, Milena de Faria Silva, Renan Carneiro Batista, and Washington Luís Ribeiro de Carvalho Segundo

An Empirical Study of Compression Strategies for Tricorder: From General-Purpose Compressors to Shannon Entropy 243
Caio Guimaraes Herrera, Hugo H. Nakamura , Isaac S. Soares , Simone R. S. Souza , and Paulo Sergio Lopes de Souza

A New Training Method for Text Classification in Mental Health: LatentGLoss 261
Korhan Sevinç

Evaluating Multi-objective Evolutionary Algorithms for Fairness-Aware Decision Tree Tuning 274
Ana Januário, Maria João Alves, and Carlos Henggeler Antunes

A Method for Automating Iterative Machine Learning Model Retraining on Event Streams in Decision Support Systems 292
Arseniy Lomakin and Aleksandr Makarov

Analysis and Improvement of the Requirements Elicitation Process Based on ISO 31000 309
Fernanda Lopes de Lima, Simone Borges Simão Monteiro, and Viviane Vasconcellos Ferreira Grubisic

General Track 6: Urban and Regional Planning

A Structured Comparative Review of Digital Structural Health Monitoring Solutions 331
Lambros Mitropoulos and Eirini Stavropoulou

Applying Transdisciplinary Experimental Methods to the Regeneration of a Former Nunnery 346
Teresa Pedretti, Pietro V. Ambrosini, Alessandro Busana, Daniele Cappelletti, and Letizia Bollini

APODigital: A Flexible Platform for Smart Surveys 362
Vinicio Bernardes Silva, Victor Guimarães, Laura Barros, Ana Julia Maria Silva, Diego Molinos, Simone Villa, and Flávio de Oliveira Silva

A Thompson Sampling Based Operator Selection in Adaptive Large Neighborhood Search for Facility Location with Customer Incompatibilities 378
Gabriel Souto, Diego Brandão, Luidi Simonetti, and Pedro Henrique González Silva

MSA Strategies for End-of-Line Automated Test Equipment with Non-replicable Measurements 396
Eduardo Santos, Eusébio Nunes, and Cristina Rodrigues

Radio Navigation Aids Network Optimization Using a Hexagonal Geospatial Indexing System 413
Ivan Ostroumov and Sergii Ivanets

Hybrid Multi-logic Reasoning Framework Over Ontologies for Decision Support in Neurodevelopmental Disorders 426
William Chaves de Souza Carvalho, Pedro Moisés de Souza, and Flávio de Oliveira Silva

The Gorizia Railway Terminal as a Strategic Hub: A Study of Logistics Dynamics in Friuli-Venezia Giulia 446
Giuseppe Borruso, Francesca Sinatra, and Saleh Samimi Dehkordi

Measuring Demographic Vulnerability and Territorial: a Composite Index Approach 456
João Lourenço Marques, Jan Wolf, and Joana Duarte

Identifying Optimal Locations for Shared e-Scooters Stations in Lexington, Kentucky 470
Eirini Stavropoulou, Dionysios Tzamakos, Lambros Mitropoulos, and Nikiforos Stamatiadis

Green and Grey Transition Data Analysis for Developing Synergies in the RECs Context 487
Lorenza Mola, Gianmarco Lorenti, Giorgia Malavasi, Lisa Panero, Amina Pereno, Maurizio Repetto, Diana Rolando, and Martina Spinelli

A Bi-Objective Model for Environmental Monitoring by Vessels and AUVs 503
Mahdi Moeini

The Effects of the Digital Transition on the Environment and Human Well-Being. From Cognitive and Evaluative Protocols to Possible Actions for Regional Planning 520
Fabrizio Battisti and Melania Perdonò

Evaluating Alternative Continuous Metaheuristic Solution Encodings for Routing Problems with Multiple Drones with Interceptions 538
Sarah Dillon and Jacomine Grobler

Shared Autonomy for an Omnidirectional Robot with SLAM and Collision Avoidance 556
Mireya Zapata, Alejandro Camino, Vannesa Vargas, and Pablo Ramos

Bi-objective Optimization of Electric Vehicle Charging Scheduling with Heterogeneous Chargers Under Time-Varying Electricity Prices 568
Imene Zaidi, Latifa Belhocine, Baptiste Buchi, and Hajar Nouinou

PHD Showcase Papers

An Operating-System Infrastructure for Embedded BDI-Based Multi-agent Systems 589
Nilson Lazarin, Carlos Pantoja, and José Viterbo

DECODE: Overcoming the Perception Bottleneck in Small Vision-Language Models for Computer Science Education 610
Martina Ianaro and Maurizio Gabbrielli

Short Papers

A Machine Learning Approach for an End-to-End Music Inpainting Pipeline 629
Gustavo Romão, Elisa T. Albergaria, Diego R. C. Dias, Marcela Alves de Almeida, Vinícius F. S. Mota, and Leonardo Chaves Dutra Rocha

A Multivariate Multi-step Deep Learning Framework for InSAR Displacement Prediction 639
Maria Maia, Luís Miguel Matos, Luís Magalhães, Joaquim Tinoco, and Steffan Davies

Computational Assessment of Greenness: Objective vs. Perceived Greenness in Urban Green Spaces 651
Amirmohammad Ghavimi and Roland Pesch

Author Index 663

General Track 5: Information Systems and Technologies

Gamifying Science Communication: An Experience Report of *Jornada Galáctica* at *Canal Ciência*

Cristiane Soares Ramos[1], Sergio Antônio Andrade de Freitas[1], Mylena Angélica Silva Farias[1], Wildemberg Sales da Silva Júnior[1], Marcus Vinicius Paiva Martins[1], and Leda Cardoso Sampson Pinto[2(✉)]

[1] Faculdade de Ciências e Tecnologias em Engenharias, Universidade de Brasília (UnB), Área Especial Indústria Proj. A, Gama, Brazil
{cristianesramos,sergiofreitas,mylena.farias,wildemberg.junior,paiva.marcus}@unb.br

[2] Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Setor de Autarquias Sul, Quadra 5, Lote 6, Bloco H, Brasília, Brazil
ledapinto@ibict.br

Abstract. The Canal Ciência Portal offers science communication content in various formats, including texts, biographies, informational materials, and interactive activities. These diverse offerings necessitate strategies to facilitate content exploration and promote recurring engagement. This paper describes the design, deployment, and initial evaluation of *Jornada Galáctica*, a gamification layer integrated into the portal, and introduces a goal-oriented measurement model to assess its development over multiple cycles. The research employs a seven-phase experience report methodology and a measurement plan based on the Goal-Question–Metric (GQM) approach. This plan integrates portal usage and navigation metrics with motivational-profile metrics. The initial measurement cycle established a baseline for the target audience's motivational profile and portal usage patterns. The subsequent cycle, conducted during the gamification layer's operation, provided preliminary engagement indicators reflecting initial adoption of the gamified experience. This study contributes a definition of gamification design and integrated evaluation process, both aligned with a goal-oriented measurement model. This alignment enables measurement goals to be directly linked to observable portal and gamification events, thereby supporting the continuous evolution of the gamification.

Keywords: Gamification · Motivational Profiling · Design Process · Octalysis Framework

1 Introduction

This study presents the design and initial evaluation of the *Jornada Galáctica* (Galactic Journey) gamification layer, which is integrated into the Canal Ciência Portal, a science communication initiative associated with a public research

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 3–20, 2027.
https://doi.org/10.1007/978-3-032-30494-0_1

institution. The evaluation employs an indicator-based analysis that combines motivation metrics derived from psychometric instruments and motivational profiles with portal usage metrics from navigation logs. These metrics are organized according to measurement goals established by the Goal–Question–Metric (GQM) approach.

From a computational perspective, the primary challenge extends beyond defining game elements to operationalizing them as observable events, software requirements, progression rules, and measurable indicators. Consequently, the intervention requires a traceable link between motivational assumptions, gamification mechanics, system instrumentation, and portal logs.

The objective of this study is to describe the process for defining and deploying *Jornada Galáctica* and to evaluate its initial implementation through measurement cycles. The first measurement cycle characterizes the target audience's motivational profile and the portal's usage patterns, establishing a baseline for comparison. The second measurement cycle, conducted with *Jornada Galáctica* in operation, yields the initial gamification-related engagement indicators.

The principal contributions of this research include: (i) a seven-phase process that integrates gamification definition with indicator-driven evaluation; (ii) a measurement model structured using the Goal–Question–Metric (GQM) approach, aligning goals, questions, and metrics with the institution's objectives; and (iii) an initial application of this model across two measurement cycles, providing baseline data and the first engagement indicators to inform the ongoing development of *Jornada Galáctica*.

The rest of this paper is organized as follows: Sect. 2 presents the theoretical background; Sect. 3 describes the methodology; Sect. 4 details the seven-phase process and the measurement plan; Sect. 5 presents the discussion; and Sect. 6 presents the conclusion.

2 Background

2.1 Gamification

Gamification is defined as the use of game design elements in non-game contexts [3]. Gamification has the potential to engage individuals, motivate action, promote learning, and help address complex problems [6].

According to Honorio *et al.* [5], one of the most widely adopted structures for gamification design is the hierarchical taxonomy proposed by Werbach and Hunter [14]. This taxonomy organizes gamification into three levels of abstraction: (i) Dynamics, which sit at the top of the hierarchy and encompass conceptual aspects of how the system evolves, including constraints, emotions (such as curiosity and competitiveness), narratives, progression, and social relationships; (ii) Mechanics, which correspond to the basic processes that engage the user and determine how the dynamics are carried out, such as cooperation, reward systems, and competition; and (iii) Components, which are the practical and visible elements in the interface, such as virtual goods, missions, teams, leaderboards, collections, points, levels, and avatars.

While Werbach and Hunter's taxonomy categorizes the structural components of gamification, it lacks a mechanism to map how these elements mobilize specific human motivational forces. For this purpose, this work adopts the Octalysis framework [2] as a lens to analyze and structure motivation based on its main core drives. Octalysis is used to analyze and apply human motivation in interactive systems. In this approach, interactive systems are structured around eight Core Drivers (CD) that influence human behavior. The CDs are:

- CD1 - Epic Meaning and Calling: People are motivated by a cause or mission they believe is greater than themselves.
- CD2 - Development and Accomplishment: The drive for progress, skill development, and achieving goals.
- CD3 - Empowerment of Creativity and Feedback: Encouraging creative freedom and providing regular feedback to guide progress.
- CD4 - Ownership and Possession: Motivation that comes from owning something and wanting to improve it or acquire more.
- CD5 - Social Influence and Relatedness: Engagement with social elements such as guidance, social acceptance, social feedback, companionship, and competition.
- CD6 - Scarcity and Impatience: The need to have something because it is rare, exclusive, or not immediately available.
- CD7 - Unpredictability and Curiosity: A drive related to the pleasure of experiencing surprises and the excitement of discovering what will happen next.
- CD8 - Loss and Avoidance: Motivation to avoid negative outcomes, including fear of loss and the desire for consistency.

To complement the Octalysis framework, this work uses Self-Determination Theory (SDT) as the conceptual basis for discussing dimensions of intrinsic motivation in the target audience, as detailed in the metrics subsection [12].

2.2 Metrics in Gamification

Evaluating the outcomes of gamification requires a multidimensional approach to establish the validity and impact of a gamified system. A model for evaluating digital interventions highlights the need to incorporate several core gamification principles [4]: (1) meaningful purpose, in which the system's goals align with user interests; (2) meaningful choice, granting users agency over how they achieve goals; (3) support for player archetypes, adapting mechanics to individual characteristics; (4) continuous feedback, communicating how actions affect outcomes; and (5) visibility, making clear what progress has been achieved and what is still needed.

The metrics used to evaluate gamification can be organized according to the type of evidence they produce and the role they play in interpreting results. The evaluation of gamified interventions cannot rely on a single type of measure, since different metrics capture distinct dimensions of the phenomenon—from observable behavior in logs to self-reported subjective states and outcomes associated

with the domain objective. Consequently, behavioral metrics provide evidence of use and interaction, while psychometric instruments support the analysis of motivational states associated with the gamified experience.

Behavioral and computational interaction metrics (collected automatically) assess objective engagement through usage logs that record events and interaction patterns with the platform. These metrics include app launches, content views, achievements, and routine logs [9]. In task-execution scenarios, researchers use click counts, the volume of generated artifacts (e.g., tags), task duration, and persistence, measured by additional attempts after meeting the minimum criterion [7,11]. In gamified educational settings, direct interactions with the gamified environment are also tracked as behavioral evidence of use [8].

To interpret this behavioral evidence in terms of motivational mechanisms, this work draws on Self-Determination Theory (SDT) [12], which explains human motivation in terms of basic psychological needs. When these needs are satisfied, SDT links this state to greater intrinsic motivation; additionally, the theory characterizes extrinsic motivation as being driven by external factors, such as rewards.

The Intrinsic Motivation Inventory (IMI) is widely used in the literature as an instrument to operationalize dimensions of intrinsic motivation, enabling the description of motivational profiles and the comparison of perceptions across groups and usage conditions. It is applied to capture subjective perceptions across seven subscales [12]:

- Interest/Enjoyment (I): Measures the interest and pleasure in engaging in an activity.
- Perceived Competence (P): Assesses the sense of competence and effectiveness in engaging in an activity.
- Effort/Importance (E): Reflects the effort dedicated and the importance attributed to engaging in an activity.
- Pressure/Tension (T): Captures the pressure or tension felt during engagement in an activity.
- Value/Usefulness (V): Evaluates the perceived utility of engaging in an activity.
- Perceived Choice (C): Measures the sense of choice and autonomy in performing an activity.
- Relatedness (R): Explores interactions and connections with others while engaging in an activity.

In this work, behavior and navigation metrics extracted from logs and Google Analytics support the evaluation of the measurement goal related to *portal use*, while psychometric metrics associated with intrinsic motivation and the motivational profile based on Octalysis support the *motivational characterization of the target audience*. To ensure the selection of metrics derived from measurement goals aligned with the portal's gamification objective, the measurement process was structured according to the Goal–Question–Metric (GQM) approach [13].

3 Methodology

This work is an experience report on the creation and initial evaluation of a gamification intervention for a science communication portal. The methodology is structured as a seven-phase process, detailed in Sect. 4, which organizes decisions related to design, implementation, instrumentation, and indicator-based evaluation. These phases establish traceability between institutional needs, gamification design choices, and the empirical evidence used in the analysis. In summary, the process comprises:

- Phase 1 – Context analysis and initial decision: characterization of the portal, the institutional purpose, and the problem to be addressed through gamification.
- Phase 2 – Motivational profile definition: definition of the target audience and its motivational profile, based on intrinsic motivation (IMI) and the Octalysis framework.
- Phase 3 – Creative ideation and co-design: specification of the narrative, missions, ranks, scoring, badges, and progression rules.
- Phase 4 – Iterative prototyping: development of the gamification layer in the portal and integration with task-execution mechanisms and interaction logging.
- Phase 5 – Deployment: release of the gamification in the production environment and activation of the logging mechanisms required for evaluation.
- Phase 6 – Evaluation, metrics, and sustainment: definition and execution of the goal-oriented measurement plan (GQM), including collection and analysis of the defined metrics.
- Phase 7 – Continuous improvement: use of measurement-cycle results to guide adjustments to mission design and instrumentation.

The portal was implemented in WordPress, the content management system for *Canal Ciência*. The gamification layer was configured using two plugins: GamiPress, responsible for points, badges, ranks, and progression rules; and Forminator, used to create quizzes and forms associated with the missions. Usage and navigation data were obtained from Google Analytics and portal logs. Motivational data were collected through questionnaires based on the Intrinsic Motivation Inventory (IMI) and the User Motivational Profile.

The evidence used in the evaluation combines behavioral metrics of portal navigation and use with self-reported motivational metrics. Usage metrics were obtained via Google Analytics, while motivational metrics were obtained through psychometric instruments (IMI) and the motivational profile structured by Octalysis (UMP). The evaluation was organized into goals, questions, and metrics according to the GQM approach [13], and conducted in cycles, including an initial baseline cycle before deploying the gamification in production and the planning of subsequent cycles with the intervention in operation.

The initial measurement cycle established a baseline for *Jornada Galáctica*. This cycle offers a snapshot of the metrics defined by the Goal-Question-Metric

(GQM) approach at the onset of the gamification operation in November 2024. Subsequent measurements can be systematically compared to this baseline to identify variations, trends, and potential effects of design and implementation adjustments. The second measurement cycle analyzed data collected between November 2024 and April 2026, following the deployment of *Jornada Galáctica*. Data from July and August 2025 were excluded from the analysis due to technical issues that prevented adequate data collection during this period.

Data collection for the evaluation of the gamification was carried out using two main sources: Google Analytics (GA) and GamiPress. GA was integrated into the WordPress platform and used to automatically record metrics related to user flow within the portal, including sessions, unique users, engagement time, session duration, bounce rate, traffic sources, and most accessed pages. These data made it possible to characterize the general navigation and interaction patterns of users within the portal.

GamiPress, in turn, was used to record events directly associated with *Jornada Galáctica*, such as journey progression, badge acquisition, score assignment, mission completion, user return, and frequency of use. Thus, while Google Analytics provided evidence about overall portal usage, GamiPress enabled the monitoring of indicators specifically related to the gamified experience.

The collected data were analyzed in aggregated form and organized according to the metrics defined in the GQM model.

4 Creating *Jornada Galáctica*

4.1 Phase 1 – Context Analysis and Initial Decision

Canal Ciência is a Brazilian government initiative dedicated to science communication, maintained by the Brazilian Institute of Information in Science and Technology (IBICT), an institution linked to the Ministry of Science, Technology, and Innovation (MCTI). IBICT works to promote competencies and to develop resources and information infrastructure in science and technology, with a focus on the production, dissemination, and integration of scientific and technological knowledge. In this context, the *Canal Ciência* Portal aims to promote access to scientific knowledge for a broad and diverse audience, using different strategies to make content accessible and to foster engagement among students, educators, and the general public. After a feasibility study and an assessment of business needs, the goal of the *Canal Ciência* Portal's gamification was defined as *increasing user engagement and encouraging participation in the proposed activities*.

With respect to ethical considerations, compliance with Brazil's General Data Protection Law was ensured [1], with an emphasis on protecting personal data and minimizing the collection of identifiable information. The main risks identified relate to the diversity of user profiles served (children, adolescents, adults, researchers, educators, and students), which may require adjustments in language, accessibility, and task complexity to avoid exclusion due to inadequate

design. Another risk concerns the possibility of progressing in the gamification without effectively engaging with the scientific content, which could reduce meaningful engagement with the scientific content. The requirement for mandatory user registration may constitute an entry barrier to *Jornada Galáctica*, especially for younger users or first-time visitors.

4.2 Phase 2 – Motivational Profile

The identification of the target audience's motivational profile was carried out following a four-phase process (Fig. 1).

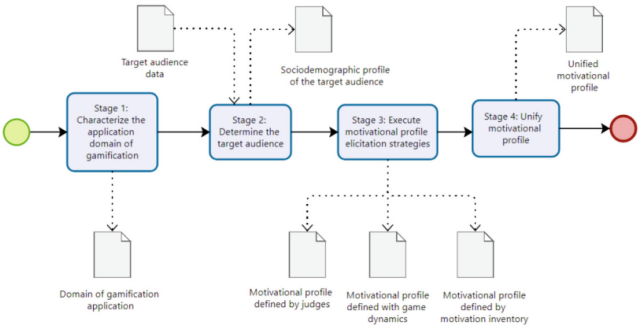


Fig. 1. Process to identify players' motivational profiles in Gamification Design [10].

In the application-domain characterization (Step 1), the domain to which the gamification would be applied was established to delimit its purpose, constraints, and target audience. In the target-audience characterization (Step 2), the primary target audience was identified, including relevant characteristics to guide the motivational-profile elicitation and, subsequently, the gamification design. The target audience comprises Brazilian students from elementary through high school who are interested in science, taking regional and socioeconomic diversity into account to ensure accessibility and adherence.

In the next phase, two strategies for eliciting the motivational profile were executed (Step 3); thus, the motivational profile was established from distinct perspectives:

– Eliciting Motivational Profile Through Judges (S1): this strategy involved selecting judges with domain knowledge, aligning gamification concepts with an emphasis on Octalysis, and then asking them to assess gamification techniques by *core drive* to estimate the target audience's profile; finally, the results from the different judges were consolidated to define the estimated profile.
  The study included 16 judges active in *Canal Ciência*, all of whom served as researchers or educators at Brazilian educational institutions ranging from

middle school through higher education. Their reported age range was 27 to 64 years; their average involvement with *Canal Ciência* was approximately 4 years (with at least two collaborators reporting more than a decade of participation), and their average educational experience was 10 years. In terms of geographic distribution, the judges represented Brazil's Central-West, North, Northeast, South, and Southeast regions. The most frequent knowledge area was Social Sciences.

– Eliciting Motivational Profile with Game Dynamics (S2): game dynamics were applied to elicit behavioral responses and preferences directly from the target audience, generating a motivational profile from this direct interaction.
Data collection involved 593 students (potential portal users) from different Brazilian states, with different levels of schooling and across age groups: children aged 5–10 years, adolescents aged 11–16 years, and young adults aged 17–24 years. A total of 581 valid responses were obtained. Data were collected with the support of teachers from different educational institutions, from elementary through high school, to obtain the greatest possible diversity of students.

Finally, the motivational profiles generated by the two strategies were unified (Step 4) into a single profile according to the procedure proposed in [10]. In S1, the judges evaluated gamification techniques associated with each Octalysis *core drive*, producing an expert-based score for each motivational dimension. In S2, the answers obtained from the game dynamics were interpreted as indicators of preference and then mapped to the corresponding *core drives*, producing a second profile based on the target audience. To combine the two profiles, the scores were organized by *core drive*, normalized to allow comparison between the expert-based and student-based results, and consolidated into a single profile. The *core drives* with the highest consolidated values were interpreted as the dominant motivational dimensions of the audience. Figure 3 was generated from this consolidated score distribution and shows the predominance of CD3, CD5, and CD6.

4.3 Phase 3 – Creative Ideation and Co-Design

The theme adopted for the portal's gamification is a space journey called *Jornada Galáctica*. It is structured as a narrative set in a future of solar-system exploration and colonization, in which players take on the role of members of multidisciplinary teams in a space race oriented toward exploration, development, and the conquest of new worlds. Narrative and mechanical progression is organized according to the Hero's Journey structure, articulating stages of player development from an initial condition to higher levels of achievement within the space-mission context.

The operationalization of the Hero's Journey in the gamification unfolds across 12 missions, in which the narrative guides the user through a set of challenges and learning experiences aligned with the central storyline. In Mission 1 (*Looking at the Earth*), the user is introduced to the problem (climate change

and pollution) and the call to action, and embarks on a space journey aboard a solar-powered spacecraft.

In Mission 1, execution is defined by an introductory letter and two initial tasks: (1) observing Earth from space and (2) identifying areas affected by pollution and deforestation, supported by an environmental monitoring panel and the solar-powered spacecraft. Completing the mission triggers rewards: the *Observador Ambiental Interplanetário* badge and the assignment of *stellar points* based on the time spent on identification (less time yields more *stellar points*), with mentorship from the ship's commander and the mascots.

Missions are defined as objectives to be fulfilled by performing actions on the portal. These actions may include, for example, accessing specific content and participating in quizzes. Records of user actions support participation and progression metrics.

Although Mission 1 is primarily structured around observation and identification tasks, CD3 is addressed through the combination of user interpretation and immediate feedback. The mission requires users to analyze visual and contextual evidence, identify areas affected by environmental degradation, and receive feedback through performance-based scoring and mission rewards. In this sense, creativity is expressed not as open-ended content production, but as situated problem interpretation and decision-making within the narrative scenario. This design choice also supports future mission expansions in which users may propose solutions, build artifacts, or personalize elements of their journey.

4.4 Phase 4 – Iterative Prototyping

In this phase, *Jornada Galáctica* was treated as a software system (a gamification layer) integrated into the portal and developed using an iterative, incremental life cycle. *Canal Ciência* is implemented in WordPress; in this work, the scope is limited to the gamification layer (*Jornada Galáctica*) integrated into the portal. First, gamification requirements and software requirements were elicited and analyzed. The gamification requirements describe the elements and rules (missions, ranks, scoring, badges, and progression). Based on these requirements, functional and non-functional requirements for *Jornada Galáctica* were derived, specifying system behavior and quality constraints. Examples of gamification requirements and software requirements are presented in Table 1 and Table 2, respectively.

In the development of gamification, *Jornada Galáctica* was implemented in WordPress using two complementary plugins. The GamiPress plugin configured and enforced progression and reward rules, including points, badges, and ranks, based on observable user actions. This configuration enabled tracking and rewarding activities such as task completion, reading portal content, and participating in quizzes, thereby updating the player's state as missions and challenges were completed. The Forminator plugin was used to create and manage forms and quizzes, facilitating the collection of user responses in knowledge-check activities across the journey's topics.

Table 1. Examples of gamification requirements

Gamification Requirements
Each mission must be configurable with respect to its rank association, while preserving the completion-bonus rule.
Jornada Galáctica must use ranks as a progression mechanism, with defined advancement criteria.
The *Looking at the Earth* mission must have an objective focused on raising awareness of the impacts of climate change and environmental preservation.
The *Looking at the Earth* mission must grant a reward composed of a specific badge and performance-based scoring.
The *Looking at the Earth* mission must use two types of scoring, with *Astropoints* tied to progression in ranks and *Energix* tied to badge-related rewards.

Table 2. Examples of functional and non-functional software requirements

Type	Software Requirements
Functional	The system must record user actions on the portal and convert them into updates to the gamification state (scoring, badges, and rank progression).
Functional	The system must provide quizzes and forms associated with missions, record responses, and link them to reward rules when applicable.
Non-functional	The system must ensure responsive behavior and navigation across different devices.

From an architectural perspective, *Jornada Galáctica* operates as a gamification layer integrated into the WordPress-based portal, following an event-driven logic. User interactions with the portal are treated as observable events that are captured, processed, and translated into updates to the gamification state. Each user action—such as accessing content, completing quizzes, submitting forms, or finishing mission tasks—generates an event recorded in the system logs. These events are interpreted by the gamification layer (configured via GamiPress), which applies predefined rules to update points, badges, ranks, and mission progression. Forminator complements this process by enabling structured data collection through quizzes and forms, which also generate events linked to gamification outcomes.

4.5 Phase 5 – Deployment

In the deployment phase, *Jornada Galáctica* was released to the portal's production environment, activating the gamification rules and the event-logging mechanisms. The gamification went live in November 2024, marking the start of operation with real users and the collection of usage data. After the release, dissemination efforts were undertaken to drive initial traffic and encourage participation, including announcements on social media and at institutional events, to increase the intervention's reach.

4.6 Phase 6 – Evaluation, Metrics, and Sustainment

To evaluate *Jornada Galáctica*, we adopted the Goal–Question–Metric (GQM) approach [13]. This approach aligns measurement goals with the organization's business needs, refines them into evaluative questions, and specifies metrics to address those questions. For *Jornada Galáctica*, two measurement goals and five questions were defined: Goal 1 (Table 3), aimed at understanding the motivational profile of portal users (questions Q1 and Q2); and Goal 2 (Table 5), aimed at understanding portal users' interests (questions Q3, Q4, and Q5). The defined metrics guide data collection to assess whether the gamification meets the established goals, identify deviations, and highlight opportunities for improvement. The relationship among goals, questions, and metrics is presented in Fig. 2, and the detailed metrics are presented in Table 4 and Table 6.

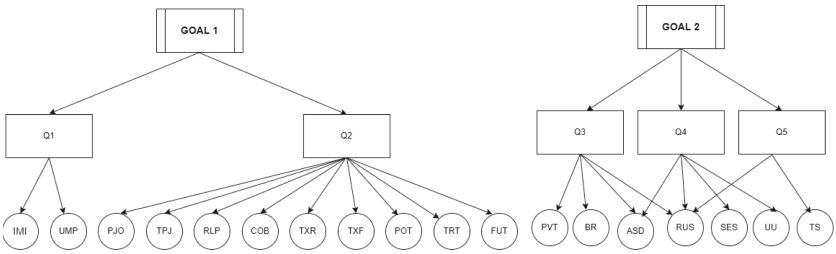


Fig. 2. Relationship between Goals, Questions and Metrics.

Table 3. Measurement Goal 1

Analyze	Gamification.
For the purpose of	Understanding.
From the viewpoint of	Gamification manager.
In the context of	Canal Ciência portal.

- Q1: What is the motivational profile of Portal users?
- Q2: How does gamification affect user engagement on the Portal?
- Q3: Which topics are of greatest interest to portal users?
- Q4: How does portal visit frequency change over time?
- Q5: Which traffic sources are most effective at attracting users interested in science to the portal?

Table 4. Measurement Goal 1 – Metrics

Code	Metric name	Description
IMI	Intrinsic Motivation Inventory	Intrinsic motivation measured through the IMI subscales.
UMP	User Motivational Profile	Intrinsic motivation is associated with Octalysis core drives.
PJO	Progression in *Jornada Galáctica*	Measures how much the user has progressed in *Jornada Galáctica*, which may indicate motivation to achieve goals.
TPJ	Time Spent in *Jornada Galáctica*	The time a user spends in *Jornada Galáctica* may indicate interest and motivation.
RLP	Long-term Retention	Measures the proportion of users who continue to use the gamified experience over a period of time, indicating sustained motivation.
COB	Achievements in *Jornada Galáctica*	Achievements earned by a user may indicate motivation to complete tasks or achieve goals.
TXR	*Jornada Galáctica* Return Rate	The proportion of users who return to *Jornada Galáctica* after a certain period of time.
TXF	Mission Completion Rate	The proportion of users who complete a mission or achieve a specified goal.
POT	Score in *Jornada Galáctica*	Measures the number of points a user accumulates.
TRT	Retention Rate	The percentage of users who return to the portal (or gamified experience) after their first visit (or over a period of time).
FUT	Frequency of Use	Measures how regularly a user returns to the portal (or *Jornada Galáctica*).

Understanding Users' Motivational Profile (Goal 1)

What is the motivational profile of Portal users? (Q1): To answer this question, GQM combines the IMI metrics (subscales) and UMP (an Octalysis *core-drive*-based profile). The most prominent *core drives* in the UMP are used as a reference to interpret which IMI subscales are most pronounced in the target

Table 5. Measurement Goal 2

Analyze	User interest.
For the purpose of	Understanding.
With respect to	Portal's usage.
From the viewpoint of	Portal manager.
In the context of	Canal Ciência portal.

Table 6. Measurement Goal 2 – Metrics

Code	Metric name	Description
PVT	Most Visited Pages by Topic	Identify the portal pages with the highest visit counts, by topic.
BR	Bounce Rate	Measure the bounce rate of science-related pages to understand whether visitors leave the portal quickly or continue browsing.
SES	Sessions	Track the total number of portal sessions over time.
UU	Unique Users (Total Users)	Monitor the number of unique users who accessed the portal during the period.
TS	Traffic Sources	Identify where visitors come from (e.g., search engines, social networks, referrals from other sites, direct traffic)
AET	Average Engagement Time per Session	Analyze how long, on average, users are actively engaged per session, which may indicate their level of interest.
RUS	Returning User Share	Measure the share of returning users out of total users, which may indicate sustained interest over time.
ASD	Average Session Duration	Evaluate the average session duration to characterize how long users spend on the portal per visit

audience (see Table 7). Thus, the motivational profile is described from two perspectives: (i) measures based on IMI subscales and (ii) motivational priorities organized by Octalysis.

Figure 3 presents the Octalysis model with the target audience's motivational profile (UMP). The results indicate that CD3 (Empowerment of Creativity & Feedback), CD5 (Social Influence & Relatedness), and CD6 (Scarcity & Impatience) are the three highest-impact *core drives*.

By relating these core drives to the IMI subscales (Table 7), we obtain a broader view of the target audience's motivation. The motivational profile points to three relevant dimensions: (i) *Interest/Enjoyment*, which reflects interest and

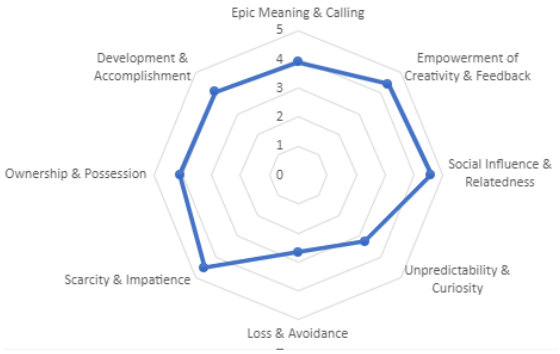


Fig. 3. Resulting motivational profile.

enjoyment in the activities; (ii) *Relatedness*, indicating that social interactions and feedback are relevant; and (iii) *Pressure/Tension*, suggesting that some level of pressure may function as a motivational factor.

Table 7. Relationship between IMI subscales and Octalysis core drives

IMI subscale	Core drives
(R) Relatedness	CD3, CD5
(C) Perceived Choice	CD2, CD4
(P) Perceived Competence	CD2, CD4
(I) Interest / Enjoyment	CD3, CD7
(T) Pressure / Tension	CD6, CD8
(V) Value / Usefulness	CD1, CD7
(E) Effort / Importance	CD2, CD8

In the second measurement cycle, the portal includes a form to collect data on the IMI and UMP instruments, which users may complete voluntarily. However, the number of responses received so far is still insufficient to generate a new motivational profile for *Jornada Galáctica* users. Therefore, the motivational profile adopted in this analysis remains as defined in the previous characterization stage, and future measurement cycles should update it as more responses become available.

How does gamification affect user engagement on the Portal? (Q2): In the first measurement cycle, the associated metrics were not collected because *Jornada Galáctica* was not yet operational. In the second measurement cycle, with gamification in operation, the metrics enabled the first indications of user engagement with the gamified experience. The PJO metric shows that most participants

(85%) remained at the first level, while 12.62% progressed to intermediate levels and 2.38% reached the final level. Engagement was therefore concentrated mainly in the initial stages of the journey.

The continuity metrics complement this interpretation: the return rate (TXR) was 21.43%, the general retention rate (TRT) was 15%, and long-term retention (RLP) was 2.38%. Among users who progressed, daily return was observed (FUT), with users remaining engaged for up to four days (TPJ). In addition, 14% of participants earned badges and received Energix (COB), while the accumulated score (POT) ranged from 50 to 5,000 stellar points, as determined by user progression. Thus, the data indicate initial adoption and signs of return, but also show that the continuity of engagement still needs improvement.

Understanding Portal Users' Interests (Goal 2)

Which topics are of greatest interest to portal users? (Q3): Analysis of the PVT metric indicates that the topics of greatest interest to users include biographies of Brazilian scientists, recent scientific output, science communication texts (*Ciência em Síntese – Science in Brief*), general information about the portal, and the history of science. This pattern indicates an interest in science content written in accessible language, as well as topics related to the history of science and scientists' contributions.

In the second measurement cycle, no significant differences were observed in the topics of greatest interest, indicating that the main patterns identified in the baseline remained stable.

How does portal visit frequency change over time? (Q4): The SES metric indicates that portal visit frequency varied throughout the first quarter of 2024, with March being the busiest period. Over the quarter, daily visits generally ranged from 200 to 600, averaging around 300, and there were nine days with more than 500 visits (peak: 601). In terms of reach, the UU metric captures the volume of unique visitors during the period, while the returning user share (RUS) was 19%, indicating that a portion of visitors returned. In addition, the bounce rate (BR) was below 1%, suggesting a low occurrence of sessions ending without further interaction. Finally, session time metrics (AET = 27 s and ASD = 1min 22 s) complement the analysis by characterizing the typical time spent per visit.

In the second measurement cycle, the SES metric recorded 21,615 sessions, while the UU metric indicated 16,689 unique users. The bounce rate remained low (BR = 0.99%), similar to the baseline, indicating that visits continued to result in some interaction with the portal. In contrast, the share of returning users was lower (RUS = 1.5%), a decrease in the relative participation of returning visitors during this cycle. The time-related metrics, however, showed a substantial increase, with an average engagement time of 2min 20 s (AET) and an average session duration of 2min 18 s (ASD). It was also observed that Monday was the weekday with the highest volume of visits.

Which traffic sources are most effective at attracting users interested in science to the portal (Q5)? In the baseline, the portal's main traffic sources (TS) were led by Google, the largest source of visits. This indicates that organic search played an important role in directing visitors to the portal. The second-largest traffic source was the organization's institutional website. Direct traffic, in third place, showed that a considerable share of visitors accessed the portal by typing the URL or using bookmarks, indicating familiarity and repeat visits.

During the second measurement cycle, the TS metric showed a change in the composition of traffic sources. Google remained the primary source of access, confirming the continued relevance of organic search; however, direct access moved from third to second place, indicating greater relative prominence for users who reached the portal directly. In addition, Linktree ranked third, while institutional pages such as gov.br ranked fourth. The set of traffic sources also became more diversified, including social networks, Wikipedia, ChatGPT, and educational platforms.

4.7 Phase 7 – Continuous Improvement

This phase is guided by the results of the second measurement cycle and by the measurement goals defined in the GQM model.

Regarding Q1, although the portal already provides a voluntary form to collect data for the IMI and UMP instruments, there are still not enough responses to update users' motivational profiles.

For Q2, the data indicate initial adoption of the gamified experience but also limited progression beyond the first level and a low mission completion rate, suggesting the need to improve instructions, feedback, progression visibility, and incentives to sustain engagement.

With respect to Q3, the stability of the topics of greatest interest and the presence of the games page among the most accessed pages suggest an opportunity to bring the journey closer to the interactive content already valued by users.

The results for Q4 show an increase in average interaction time, but a lower proportion of returning users, indicating the need for strategies that encourage return visits, such as calls to progress to the next missions and indicators of pending progress.

Finally, Q5 shows that, although Google remains the main traffic source, direct access has gained relevance, and access origins have become more diversified, reinforcing the use of channels such as social networks, institutional pages, Linktree, and educational platforms to direct users to specific missions and to strengthen the continuity of the gamified experience.

5 Discussion

The analysis of the results should consider the complementary role of the measurement cycles. The first cycle established a baseline for the motivational profile

and portal usage patterns, while the second cycle provided the first indicators of *Jornada Galáctica* in operation. User progression was concentrated mainly at the first level, with only a small proportion reaching the final level, suggesting stronger engagement in the initial stages of the journey. This limited progression may be related to the requirement to read specific portal texts in order to advance, a condition that may represent a barrier for some users.

The interpretation of these findings also depends on updating the motivational profile. The reference profile, characterized by the predominance of CD3, CD5, and CD6, has not yet been reassessed using data from actual *Jornada Galáctica* users, as the voluntary form based on the IMI and UMP has not yet received enough responses. Thus, it remains open whether the observed behavior is mainly due to limitations in implementation, differences in users' motivational profiles, or a combination of both.

Overall, the results indicate initial adoption of gamification but also difficulties sustaining engagement throughout the journey. Although return visits were less frequent, users who accessed the portal tended to spend more time interacting with its content.

Changes in traffic sources and the stability of the most accessed topics should be analyzed, taking into account dissemination actions, institutional events, content updates, seasonality, and technical limitations in data collection. These findings reinforce the need to adjust the gamified experience and to establish new measurement cycles to evaluate its evolution.

6 Conclusion

This work presented an experience report on the design, deployment, and initial evaluation of the *Jornada Galáctica* gamification layer integrated into the *Canal Ciência* portal. The GQM approach was used to structure measurement goals, questions, and metrics, combining behavioral evidence of portal use with motivational evidence obtained from psychometric instruments and the motivational profile structured by Octalysis.

The results of the first measurement cycle were used as a baseline to characterize the target audience's motivational profile and portal usage patterns before gamification data became available. The second measurement cycle provided the first indicators of user engagement with the gamified experience. Although *Jornada Galáctica* achieved initial adoption, sustaining engagement throughout the journey remains a challenge.

As future work, the study plans to refine the gamified experience and update the motivational profile with data from actual users and additional measurement cycles.

Beyond the specific case of *Canal Ciência*, this study contributes a traceable evaluation process that integrates users' motivational profiles, gamification design decisions, interaction-logging mechanisms, and observable engagement indicators.

References

1. Brazil: Federal law no. 13709 of 14 august 2018 (general data protection law – lgpd). Presidency of the Republic, Federal Government of Brazil (Planalto) (2018). https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, Accessed 19 Mar 2026
2. Chou, Y.K.: Actionable gamification: Beyond points, badges, and leaderboards, Createspace Independent Publishing Platform (2015)
3. Deterding, S., Dixon, D., Khaled, R., Nacke, L.E.: From game design elements to gamefulness: defining "gamification". In: International Conference on Entertainment and Media in the Ubiquitous Era (2011). https://api.semanticscholar.org/CorpusID:207193782
4. Floryan, M., Chow, P.I., Schueller, S.M., Ritterband, L.M.: The model of gamification principles for digital health interventions: evaluation of validity and potential utility. J. Med. Internet Res. **22**(6), e16506 (2020). https://doi.org/10.2196/16506
5. Honório, J., B. Moura, J.A., Brito, P., Menezes, T., Barros, M.: Gamification in Location-based M-Learning: Students' Perceptions of Game Elements:. In: Proceedings of the 10th International Conference on Computer Supported Education, pp. 489–496. SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal (2018). https://doi.org/10.5220/0006716404890496
6. Kapp, K.M.: The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education. Pfeiffer, San Francisco, Calif (2012)
7. Na, K., Han, K.: How leaderboard positions shape our motivation: the impact of competence satisfaction and competence frustration on motivation in a gamified crowdsourcing task. Internet Res. **33**(7), 1–18 (2023). https://doi.org/10.1108/INTR-12-2021-0897
8. Puig, A., Rodríguez, I., Rodríguez, Á., Gallego, I.: Evaluating learner engagement with gamification in online courses. Appl. Sci. **13**(3), 1535 (2023). https://doi.org/10.3390/app13031535
9. Rajani, N.B., Bustamante, L., Weth, D., Romo, L., Mastellos, N., Filippidis, F.T.: Engagement With Gamification Elements in a Smoking Cessation App and Short-term Smoking Abstinence: Quantitative Assessment. JMIR Serious Games **11**, e39975 (2023). https://doi.org/10.2196/39975
10. Ramos, C.S., Farias, M.A.S., De Freitas, S.A.A., Martins, M.V.P., Alves, J.M., Pinto, L.C.S.: A Process to identify players' motivational profiles for designing a gamification project. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., Rocha, A.M.A.C., Faginas Lago, M.N. (eds.) Computational Science and Its Applications – ICCSA 2024, vol. 14814, pp. 49–67. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-64608-9_4
11. Rose, J., O'Meara, J., Gerhardt, T., Williams, M.: Gamification: Using elements of video games to improve engagement in an undergraduate physics class. Phys. Edu. **51**, 055007 (2016). https://doi.org/10.1088/0031-9120/51/5/055007
12. Ryan, R., Deci, E.: Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness, Guilford Publications (2017)
13. van Solingen, R., Berghout, E., Berghout, E.: The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development. The McGraw-Hill Companies, London (1999)
14. Werbach, K., Hunter, D.: For the Win: How Game Thinking Can Revolutionize Your Business. Wharton Digital Press, Philadelphia, EBL-Schweitzer (2012)

Mitigating Membership Inference Attacks on Knowledge Graphs

Jnaini Abdellah(✉)

Department of Computer Science, Mohamed First University, Oujda 60000, Morocco
abdellah.jnaini@ump.ac.ma

Abstract. Knowledge Graphs (KGs) are structured representations that model entities and their relationships in a connected form. By leveraging their capabilities in the link prediction task and node classification on graphs, They have been successfully used in a variety of fields, including drug discovery, community detection, and location sharing applications. The usage of KGs in privacy-sensitive contexts has been made easier by these potent uses and the wide availability of graphs in various domains. Unfortunately, KGs are susceptible to well-defined attacks that could reveal private information. Membership inference attacks, which enable the attacker to determine if a certain sample was included in the training dataset, are our primary area of interest. In order to avoid this, we present three LLM-guided defense techniques that are implemented at the posterior level: knowledge distillation, secure aggregation, and posterior encoding with noise. Our suggested methods preserve the model's performance on the link prediction task while successfully reducing MIA accuracy. Our results provide insights into striking a balance between privacy preservation and predictive performance, and they have been confirmed through extensive tests on popular KG designs.

1 Introduction

The recent accomplishments of Deep Neural Networks (DNNs) have revolutionized the field of Machine Learning (ML), especially in machine translation [1], speech recognition [2], and object detection [3] tasks. This revolution is ascribed to the progress of Deep Learning (DL) architectures, such as Convolutional Neural Networks (CNNs) [4], Recurrent Neural Networks (RNNs) [5], and Autoencoders [6]. The input representation is obtained concurrently with the downstream task within the deep learning architecture. The process of collaborative learning requires no manual feature development because it operates differently from traditional machine learning methods. The processing power improvements have enabled better performance of neural architectures through the development of advanced and more intricate systems. The extensive training datasets which researchers now have access to, enable better learning and generalization capabilities, which lead to improved performance across different tasks. The networks successfully extract hidden features from Euclidean data, which includes

© The Author(s), under exclusive license to Springer Nature Switzerland AG 202
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 21–37, 2027.
https://doi.org/10.1007/978-3-032-30494-0_2

photos and text and videos, thus enhancing their performance across multiple use cases. Convolutional Neural Networks (CNNs) have proved essential in understanding and utilizing the attributes of picture data [7].

Nonetheless, the limitations of conventional deep learning architectures become evident when dealing with graph-structured data, as machine learning models progress. Graphs are ubiquitous in our daily lives and represent a diverse array of information, including chemical structures in chemistry and social interactions. The standard assumptions of machine learning, including instance independence, are not maintained in the graph domain, resulting in considerable difficulties in implementing conventional deep learning architectures on graph data [8].

Knowledge graphs (KGs) effectively handle various tasks by leveraging intricate graph structures and corresponding node properties. Moreover, KGs demonstrate versatility across several scales of graph analysis. At the node level, they utilize message passing and graph convolutions to generate representative node embeddings, typically employing softmax or multilayer perceptron layers for final predictions [9]. The system uses edge-level tasks which depend on node representations and similarity metrics and neural architectures to perform link prediction and edge classification tasks [10]. Graph-level analysis combines pooling and readout operations to create complete graph representations which can be used in classification tasks [11].

According to security assessments, KGs demonstrate effective performance. However, these systems present major data privacy problems. The systems face security threats because their message passing features. They allow to attackers to use membership inference, attribute inference and property inference attacks [12].

Therefore, there is a need to implement sophisticated security practices and protection mechanisms due to the aforementioned risks that can affect operations handling sensitive information.

In this paper, we analyze membership inference attacks targeting knowledge graph models used in the link prediction task. A membership inference attack constitutes one of the most dangerous attacks as it attempts to reveal whether a certain point was used for training the KG-powered machine learning model. They exploit different characteristics of the model (e.g., confidence levels and variance) to extract sensitive information from the model about the training data. For instance, in the context of knowledge graphs, it implies identifying the triple (h, r, t) contained in the training graph and endangering sensitive information.

A number of defense strategies for protecting KGs against the aforementioned risks (specifically membership inference attacks) have been introduced through various approaches. These include the utilization of probability encoding to obfuscate the connections between classes and graph perturbations to deliberately alter the topology and nodes' features. Although these defense strategies have a significant potential to reduce the risk of being attacked, they have

one critical drawback. They always imply the need to strike a balance between increased security and model performance since there is no way around the issue.

To deal with the challenge of protecting KGs against membership inference attacks, we introduce three LLM-guided techniques of posterior encoding that prevent an attacker from detecting the membership of a certain item in the training set. Our methodologies are intended to make membership untraceable within the output of our model:

- **Posterior Encoding with Noise**. This approach consists in intentionally disrupting the posterior probabilities through adding noise to them, making it more difficult for the attackers to distinguish between members and non-members. We introduce a parameter (ϵ) to regulate the amount of introduced noise.
- **Knowledge Distillation**. Using a teacher KG model that generates soft labels for the student model. The soft labels help achieve generalization of outputs and avoid overfitting to the training data, which in turn makes the membership untraceable.
- **Secure Aggregation**. Training several independent KG models and aggregating their outputs. The combined output will undergo further disruptions using noise.

All three of the proposed methods ensure that the index of the maximum posterior value stays at its peak. Together, all of the above approaches obfuscate the membership status of data entities while preserving link prediction accuracy. Based on the obtained results, we claim that these techniques successfully reduce the impact of MIA attacks on various KG structures and corpora.

All three of the proposed methods ensure that the index of the maximum posterior value stays at its peak. Together, all of the above approaches obfuscate the membership status of data entities while preserving link prediction accuracy. Based on the obtained results, we claim that these techniques successfully reduce the impact of MIA attacks on various KG structures and corpora.

This document has the following structure: A literature review, covering prior works on the topic in Sect. 2. The proposed method is presented in detail in Sect. 3, with an extensive experimental section, describing the techniques and parameters used, presented in Sect. 5. Next, our results are discussed in Sect. 6. This document concludes with the conclusion and potential future research questions outlined in Sect. 7.

2 Related Works

The studies of Membership Inference Attacks (MIA) in graph-based models have made substantial progress. Previous researches [13,14] have unveiled vulnerabilities in KGE tasks and paved the way for future privacy analyses. One of the major breakthroughs in this realm was brought about by [15] due to their attack mechanism for KG MIA based on shadow dataset/models generation. Further advancements in this sphere were introduced by [16,17], who enriched the attack

model framework by introducing multi-hop confidence vectors in its training procedure, thus improving inference accuracy. Additional improvements have been achieved in terms of better understanding of vulnerabilities in KGE models. Disparate vulnerability patterns in link inference attacks have been noted by [18,19]. As a response to these vulnerabilities, defensive measures for protecting KGE-based triple prediction from MIA have been designed using three major approaches. Encoding-based measures alter probability distributions of outputs to increase resistance towards Model Inversion Attacks via multiple mechanisms. One-hot encoding [20] uses binary vector representations to retain only the highest probability of all classes in each posterior vector while keeping its representation similar for different triples. The ReLU encoding [21] keeps only the probability value of the highest class. Then sets other values to zero. The Posterior shuffling method [22] retains the highest probability and permute indices of all other classes. Finally, the top-k selection restricts the output to classes with the highest probabilities. Perturbation-based techniques involve adding controlled perturbations to KG embeddings, which often rely on using knowledge graph auto-encoders to manipulate embeddings in such a way to keep triples structurally consistent. In turn, regularization focuses on mitigating overfitting using Dropout, in which case certain neurons get deactivated randomly in each iteration in order to minimize the influence of particular embedding dimensions. Based on the aforementioned survey, defensive strategies to KGE model attacks may be classified into three major categories: (i) Preprocessing of data and graphs before the training stage. (ii) Training-stage defense measures, including adversarial training, model architecture changes, differential privacy (DP) and federated learning (FL). (iii) Post-training/post-processing defense measures such as detection and output perturbation. Our proposed solution of LLM-guided posterior encoding definitely falls under the category of post-training/post-processing defense since our approach operates on posterior vectors generated during the training process. At the same time, our approach introduces a novel aspect to the currently known taxonomy of defense techniques. Instead of using pre-existing perturbation schemes, our technique dynamically encodes posterior vectors using transformer optimization while keeping the index of highest likelihood, thus ensuring high utility of the model without revealing membership information. Bringing a novel method inside the post-training segment of the mentioned taxonomy. Compared to DP and FL approaches belonging to the second category, our approach bypasses the disadvantages listed in the survey such as decreasing utility by introducing excessive noise (DP) and high communication costs with gradient leakage (FL). However, unlike other methods requiring retraining or distributed setup, our approach requires neither, relying solely on low-dimensional posterior vectors (the dimension of which depends on the number of classes in a dataset). That makes it lightweight, flexible and computationally efficient in agreement with the survey requirement of methodologies that reconcile privacy, reliability and environmental sustainability of robust KGEs. Encoding-based approaches reduce the amount of information revealed after learning, although they apply strict and inaccurate transformations. Per-

turbation approaches increase the robustness of the KGE but require additional steps to re-train it with knowledge graph autoencoder. That also makes them computationally expensive. The application of regularization techniques, such as dropout, helps prevent overfitting in the learning process. However, they provide indirect protection and fail to withstand robust membership inference attacks. These defense techniques can be highly effective at reducing model vulnerability to MIA, though they influence negatively its performance. Trade-offs between model security and utility lie behind the proposed methodology that seeks to find the optimal balance between resistance to MIA and effective prediction of relations within a knowledge graph.

3 The Proposed Approach

3.1 Attack Procedure

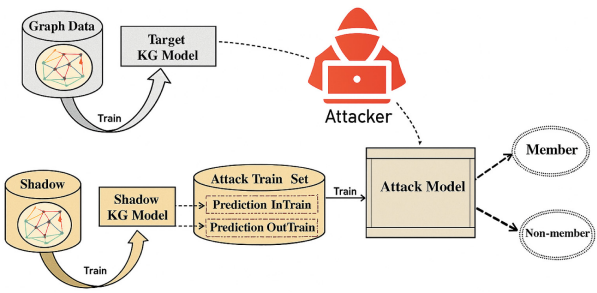


Fig. 1. Analysis of MIAs on Knowledge Graph models.

Membership Inference attacks represent a binary classification task whose purpose is to identify whether a certain triple $\tau = (h, r, t)$ defined by embeddings of entities and relations is contained in the training set of a certain KGE. First, let us name our attacker's classifier as $\mathcal{A}$. Second, we will use the accuracy metric to evaluate the attacker's abilities to infer the inclusion of a triple into a training dataset of the target KGE. Accuracy is defined as the ratio of the predicted memberships of triples to the actual number of their presence in the training dataset. In order to train the attacker's classifier $\mathcal{A}$, we need to develop a labelled dataset containing the ground truth information. For this purpose, we train a shadow KGE model that behaves similarly to the target model. It means that the shadow model uses the same architecture and hyperparameters as the target model (e.g. TransE, TransH, DistMult, ComplEx). Here we propose three phases for creating the attack: // 1/Training shadow model requires the attacker to use the same architecture as the target model. The shadow model needs to be trained on a dataset representing knowledge graph, similar to the one used for the target KGE. To this end, the attacker should have access to two datasets:

D^{train}_{shadow} – the training one and D^{test}_{shadow} – the test one.
2/Binary classifier $\mathcal{A}$ building requires gathering the posteriors from the trained shadow model. Posterior of each triple from D^{train}_{shadow} is considered to be a member, while posterior from D^{test}_{shadow} – a non-member. Labels of this kind represent ground truth information. Next, the gathered posteriors together with labels will be used to train the attacker's classifier $\mathcal{A}$.
3/Membership inference is the step when the attacker wants to know whether a specific triple belongs to the training set of the target KGE model. To do so, the attacker queries the model about the posterior or plausibility score of the candidate triple. Having received the posteriors, the attacker feeds them into classifier $\mathcal{A}$ (Fig. 1).

For enhancing MIA protection for a certain type of datasets, one can use an adaptive classifier. Thus, we select the best performing ML algorithm of classification from multiple candidates (attack models $\mathcal{A}$), for example, MLP [23], XGB [24], and Random Forest [25].

4 Proposed Defense Strategies

In this paper, we propose three strategies for defending the Knowledge Graph (KG) against MIAs by encoding posterior probability using an LLM. Our proposed techniques are intended to obfuscate attacker's capability to infer membership status by changing class posteriors while ensuring that classification accuracy remains unchanged. In our defense technique, we adopt an LLM architecture, not as an auxiliary tool, but as a module inside the KG model that is applied to the KG model's outputs (logits or posteriors) to generate posterior distributions dynamically. The role of the LLM model is to find a flexible mapping between KG model's outputs and new posteriors. By adopting LLM architecture for manipulating posteriors, we create adaptive, learnable and context-dependent techniques, which cannot be achieved in static posterior perturbation.

Knowledge distillation involves using the knowledge learned by the teacher model in the training of a student model through the alignment of output distributions. In this case, soft output is obtained by tempering the logit outputs with a temperature parameter τ. Static tempering cannot handle variability in overconfidence, class imbalance or local KG graph structure.

Our defensive strategy uses the teacher model trained on the dataset, and we soften the outputs with a temperature parameter τ to yield "soft labels", which will serve as input to the training of the student. Using soft labels as opposed to the hard classifications makes the student model learn the underlying distribution of data, thereby avoiding overfitting to particular training examples.

To improve this process, we introduce a transformer model g_ϕ using LLM architecture, which takes the logits $\mathbf{z}_v^{(t)} \in \mathbb{R}^C$ of the teacher for each triple v and computes the softened posteriors $\tilde{\mathbf{p}}_v = g_\phi(\mathbf{z}_v^{(t)})$. The objective function used to train the student model is KL-divergence with added entropy minimization. Using this architecture allows for dynamic and triple-dependent tempering, protecting KG nodes, which might be vulnerable to attacks based on their posterior

distributions. Noise is introduced to the posteriors to make inference about the triple's membership even harder. Introducing noise hinders the attacker to make any conclusion about membership status purely based on posteriors. The noise generated by a noise function parameterized by ϵ is added to all posteriors making it impossible for the attacker to identify the predicted probability value.

Posterior Encoding using LLM-Guided Secure Aggregation

In the strategy of secure aggregation, outputs of several independent KG models are used to mask membership signals. Normally, aggregation of posteriors is done using a simple average. This strategy considers all examples identically ignoring the degree of agreement between the models, vulnerability of a particular triple and imbalance in classes. As part of our defense scheme, we train several independent KG models, wherein each produces class probabilities for every entity or triple. We use averaging for pooling the predictions, thus preventing the attack that relies on the results of a particular model.

Instead of simple averaging, we propose a model g_ϕ inspired by a large language model (LLM), which accepts as an input a stacked vector of logits from all models: $\mathbf{Z}_v = [\mathbf{z}_v^{(1)}, ..., \mathbf{z}_v^{(K)}]$. Suppose $\mathbf{Z}_v \in \mathbb{R}^{K \times C}$ yields an aggregated posterior:

$$\tilde{\mathbf{p}}_v = \text{softmax}(g_\phi(\mathbf{Z}_v) + \boldsymbol{\eta}), \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \sigma^2 I)$$

The transformer utilizes disagreement between different models' outputs and offers adaptive smoothing to obscure reliable predictions, which might serve as membership evidence. We train the module with the help of:

$$\mathcal{L}_{\text{agg}} = \text{CE}(\tilde{\mathbf{p}}_v, y_v) - \lambda \cdot \mathcal{H}(\tilde{\mathbf{p}}_v)$$

We inject noise into the aggregated posterior predictions to make them even harder to guess, making the probability of an accurate inference lower. The prediction in question is made based on the largest probability in the perturbed posterior, normalized for the latter to provide valid probabilities. As a result, adversaries will not be able to use the aggregated posterior to infer membership.

Posterior Encoding with LLM-Guided Noise Injection

Noise injection reduces the likelihood of membership inference attacks by adding uncertainty to output predictions. Conventional approaches use either Gaussian or Laplacian noise.

However, the same amount of noise neglects context, class confidence, and vulnerability to membership inference attacks for KG components.

As part of our defense scheme, we only train one single KG model for every dataset, wherein noise is injected to the posterior of each triple during the inference phase. The noise, sampled according to a noise function ϵ parameter, is then applied to each posterior probability. For our LLM-based method, the transformer model g_ϕ learns to generate context-dependent noise. The above noise scheme guarantees top-1 accuracy of the classifier while achieving minimum entropy.

Upon adding noise, we perform normalization of posteriors to ensure that they correspond to probability distributions. To avoid large perturbations in the

projected class, we swap the maximum probability class with any other class that may be affected. When the effect of the added noise is negligible, we retain the original projection and choose the maximum perturbed probability.

The discussed techniques provide robust protection from the membership inference attacks, while affecting only marginally the performance of the KG model in link predictions.

4.1 Architecture of the LLM Model

LLM (Generative Pre-trained Transformer) employs the architecture of a decoder-only transformer model. Initially, the architecture was proposed for autoregressive language modeling tasks. In such models, the input sequences are transformed into vectors and are processed using attention mechanisms to identify the contexts between individual elements. The model is extremely efficient in posterior transformation tasks and, therefore, is a perfect choice to implement privacy preservation measures for knowledge graphs (KG).

The inputs to the LLM model are distinct depending on the type of applied technique. For knowledge distillation, the input will consist of one posterior vector $X = [\mathbf{p}_v^{(t)}]$ which represents a soft label generated by the teacher model. In case of secure aggregation, the input will contain multiple posterior vectors belonging to an ensemble of classifiers, i.e., $X = [\mathbf{p}_v^{(1)}, \ldots, \mathbf{p}_v^{(K)}]$. Finally, for methods based on noise injection, the input will be formed by a single posterior vector $X = [\mathbf{p}_v]$, which will then be disturbed by the model. In all cases, we may use optional positional encodings $p_t \in \mathbb{R}^d$. The self-attention mechanism forms the backbone of the LLM model architecture. At layer l, the inputs are represented by query, key, and value projections through the use of learned weight matrices. Here, h_{l-1} is the hidden state at the previous layer, and T denotes the number of tokens in the sequence. The use of self-attention enables the model to selectively focus on relevant parts of the input sequence, thereby enhancing the ability to capture complex dependencies.

Next, following the attention block, each layer has a position-wise feed-forward network (FFN), where the input is further processed through a two-layer perceptron with ReLU activations:

$$\mathrm{FFN}(x) = \mathrm{ReLU}(xW_1 + b_1)W_2 + b_2$$

After processing by the attention and feed-forward sub-layers, the results are combined via skip connections and normalized with layer normalization to give the output of the layer where W_1, W_2 are the weights of the feed-forward layers, while b_1, b_2 are biases.

Depending on the chosen defense mechanism, the output layer also changes accordingly. For knowledge distillation, the final hidden representation is mapped to class probabilities using softmax function.

Here, W_{out} refers to the output projection layer and h_L is the final hidden representation produced by the last layer L. On the other hand, in secure aggregation, a special classification token (`[CLS]`) is appended to the input sequence,

whose final embedding is used to produce the output. When it comes to noise injection, the model learns to predict a perturbation vector that is then used to add noise to the posterior. There are several advantages associated with the use of the LLM for this problem. First, it enables the learning of complicated nonlinear transformations of logits and posterior distributions, which helps to obfuscate membership information. Additionally, depending on the KG elements' risk profile, the model will be able to adaptively modify such transformations. Finally, because of the end-to-end nature of the model, it can directly optimize the trade-off between accuracy and privacy protection.

5 Experiments

5.1 Datasets

Experimental studies have been conducted on five datasets including four knowledge graphs (KGs): WN18RR, FB15K237, NELL-995, and two proprietary KGs (DDB14 and FinKG). While publicly available KGs were used for training, proprietary KGs helped us evaluate the effectiveness of the proposed framework in protecting against any privacy leakage.

The choice of these KGs was made based on the varied sizes and relational complexities. For example, WN18RR and DDB14 can be defined as small-scale KGs, whereas FB15K237 can be referred to as medium KGs and NELL-995 and FinKG are large-scale KGs. In addition, we believe that this difference in scales may assist us in evaluating the performance of our approach on different KG scales.

For creating the target and shadow KG models, each KG was then further split into two equal-sized subsets having similar distributions in entities, relations, and triples. The first subset is taken to form the target model data that will be noted as D_{Target}. Next, this D_{Target} data will be partitioned into two datasets that will be called D^{train}_{target} and D^{test}_{target}. Similarly, the second subset will act as shadow model data denoted by D_{Shadow}. This D_{Shadow} dataset will be then split into two datasets, called D^{train}_{shadow} and D^{test}_{shadow}.

5.2 Proposed Approach

Our choice of LLM was GPT-4 which was used for the guidance in the posterior encoding of the knowledge graph (KG) models. The LLM guidance of the posterior vector encoding was introduced to mitigate the effect of MIA attacks for the knowledge graph models. We apply three approaches based on LLM posterior encoding to improve the KG models' ability to protect itself from membership inference attacks. For our research experiments, we use the two-layer architecture for the following KG models: TransE, TransH, DistMult, and ComplEx. The models were trained for 200 epochs in each case. It is important to highlight the necessity to know the experiment setting in detail when reproducing our experiment in the future for mitigating the attacks on KG models. Our approach does not include any text data or graphs since the encoder works with the

posterior vectors outputted by the last KG layer. Every posterior represents the probability distribution of the triple score, and the dimensions of the vector are equal to the number of the relation types in the dataset.

Transformer modules take posterior vectors to encode them into the context-aware perturbations, that in turn can be used in post-hoc perturbation. This approach makes the defense method lightweight (≈12M parameters) but effective thanks to the expressive nature of the transformers.

5.3 Experimental Setup

The setting of the experiment was made carefully in order to enhance the effectiveness and speed. The system was configured to use TPU (32 GB RAM) acceleration for the computations. The latter can be achieved due to TensorFlow (version 2.13.0). TensorFlow is an open-source library developed by Google and allowing to use available TPUs in the automatic way. Increased RAM usage allows dealing with large-scale datasets and computations efficiently. Our main programming language was Python (version 3.10.12). Our experiments used various libraries.

Also, several special libraries required for knowledge graph (KG) handling such as OpenKE, PyTorch, and additional libraries necessary for KG training and scoring were included in order to handle the complexity of the KG models' requirements. Special KG manipulation is vital for KG models and allows performing scoring, triples' negative sampling and other processes related to triples handling.

5.4 Metrics

In order to estimate the performance of the attack and the defense methods, we provide three well-known metrics: **Accuracy**, **Precision**, and **F1 score**.

Accuracy is the measure of the ratio of correct membership decisions. Precision describes the proportion of true positives. Meanwhile, F1 score is a harmonic mean of Precision and Recall, meaning the balance of these two measures.

Considering additional criteria for evaluation of the attack/defense methods, the research became more comprehensive and informative.

6 Results and Discussion

6.1 Mitigation of Membership Inference Attacks via Three LLMs Based Defenses

In this section, we shall delve into the empirical observations gathered using the three LLM-guided defense mechanisms mentioned above: LLM* noise injection, LLM** secure aggregation, and LLM*** knowledge distillation. The experiment involves five benchmark datasets (WN18RR, FB15K237, NELL-995, DDB14, and FinKG) and four leading knowledge graph embedding models (TransE,

TransH, ComplEx, and DistMult). The y-axis shows the accuracy of the membership inference attack (MIA), which is inversely proportional to privacy safety, such that the smaller the number, the higher the degree of security achieved. For random guessing, the MIA accuracy score is set to 50%, and any defensive measures obtaining figures near or below this threshold can be considered effective. All three LLM-guided strategies reduce the accuracy of the MIA, compared to the unaltered baseline (approximately 95–97%). Therefore, the posterior encoding approach controlled by the LLM mechanism is proven efficient. Out of all the techniques, the LLM guided by knowledge distillation performs best, providing uniform protection for various models and datasets, and ensuring that MIA accuracy stays under 47% on average. The lowest values are recorded at 41.8% for TransH on WN18RR and 41% for DistMult on FB15K237, illustrating high uncertainty in classification decisions without affecting overall precision. The performance is attributed to the LLM's ability to flexibly weaken posteriors while retaining top predictions and increasing their entropy (Fig. 2).

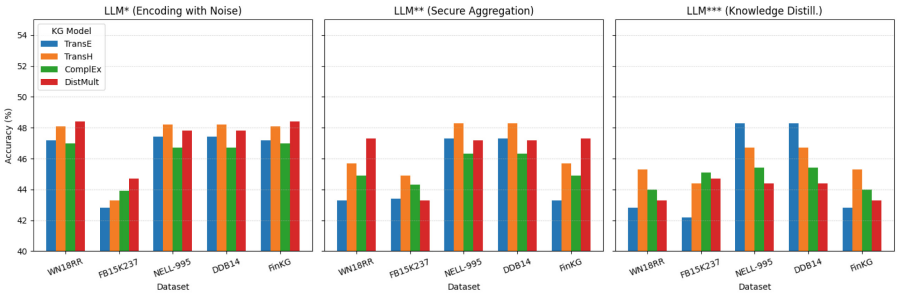


Fig. 2. Attack model accuracy against different KG models after applying LLMs defenses approaches.

Meanwhile, the LLM guided by secure aggregation implements an encryption algorithm in aggregating several knowledge graph embedding posteriors, resulting in MIA accuracy levels between 42% and 46%. The transformer-based aggregation model introduces disagreements among inter-model decisions, reducing overconfidence in predictions. Therefore, its efficacy is demonstrated particularly in NELL-995 (for example, at 44% for TransE and 45.0% for ComplEx). Thus, it is validated that multi-model reasoning controlled by a large language model can achieve substantial privacy protection.

In summary, LLMs that employ guided noise injection into posteriors achieve slightly higher MIA accuracies (around 44–49%) while offering considerably enhanced privacy protection compared to the baseline level. This slightly worse performance can be explained by the possible inability to create sufficient structured uncertainty without compromising utility through noise injection. However, this methodological framework remains attractive because of its simplicity and node/class specificity of the LLM module.

Table 1. Accuracy of membership inference attack after employing different defense approaches.

Dataset	Defenses	TransE	TransH	ComplEx	DistMult
WN18RR	MIA	89.63	89.22	93.56	93.01
	Shuffle ENC.	52.90	58.30	58.30	59.70
	Dropout	51.70	54.80	53.60	52.10
	One Hot ENC.	50.00	50.00	50.00	50.00
	RELU	63.40	58.10	60.10	58.90
	K TOP	67.30	72.00	55.20	57.40
	VAE	57.30	57.50	55.20	55.70
	Graph per 1	55.10	55.70	54.70	56.20
	Graph per 2	54.90	55.60	55.20	54.70
	LLM*	47.20	48.10	47.00	48.40
	LLM**	43.30	45.70	44.90	47.30
	LLM***	42.80	45.30	44.00	43.30
FB15K237	MIA	82.87	84.41	72.65	67.13
	Shuffle ENC.	58.10	53.50	51.00	56.20
	Dropout	61.80	57.70	53.40	53.70
	One Hot ENC.	50.00	50.00	50.00	50.00
	RELU	57.00	58.20	57.90	58.30
	K TOP	67.50	66.40	67.90	68.20
	VAE	57.10	58.70	57.70	56.40
	Graph per 1	56.70	55.30	56.60	53.30
	Graph per 2	53.10	55.20	54.30	54.90
	LLM*	42.80	43.30	43.90	44.70
	LLM**	43.40	44.90	44.30	43.30
	LLM***	42.20	44.40	45.10	44.70
NELL-995	MIA	72.45	84.23	84.04	81.78
	Shuffle ENC.	56.90	59.40	59.00	59.70
	Dropout	50.00	50.00	50.00	50.00
	One Hot ENC.	50.00	50.00	50.00	50.00
	RELU	67.90	68.00	67.10	67.50
	K TOP	63.30	61.60	60.90	59.40
	VAE	56.10	57.10	59.10	57.50
	Graph per 1	54.80	57.00	57.70	57.30
	Graph per 2	54.30	56.70	56.90	57.00
	LLM*	47.40	48.20	46.70	47.80
	LLM**	47.30	48.30	46.30	47.20
	LLM***	48.30	46.70	45.40	44.40

(*continued*)

Table 1. (*continued*)

Dataset	Defenses	TransE	TransH	ComplEx	DistMult
DDB14	MIA	94.76	95.38	94.00	92.60
	Shuffle ENC.	63.90	59.40	58.90	57.70
	Dropout	56.90	59.70	59.40	59.70
	One Hot ENC.	50.00	50.00	50.00	50.00
	RELU	63.30	61.60	60.90	59.40
	K TOP	63.30	61.60	60.90	59.40
	VAE	56.10	57.10	59.10	57.50
	Graph per 1	54.80	57.00	57.70	57.30
	Graph per 2	54.30	56.70	56.90	57.00
	LLM*	47.40	48.20	46.70	47.80
	LLM**	47.30	48.30	46.30	47.20
	LLM***	48.30	46.70	45.40	44.40
FinKG	MIA	99.80	99.98	86.66	83.58
	Shuffle ENC.	52.90	58.30	58.30	59.70
	Dropout	51.70	54.80	53.60	52.10
	One Hot ENC.	50.00	50.00	50.00	50.00
	RELU	63.40	58.10	60.10	58.90
	K TOP	67.30	72.00	55.20	57.40
	VAE	57.30	57.50	55.20	55.70
	Graph per 1	55.10	55.70	54.70	56.20
	Graph per 2	54.90	55.60	55.20	54.70
	LLM*	47.20	48.10	47.00	48.40
	LLM**	43.30	45.70	44.90	47.30
	LLM***	42.80	45.30	44.00	43.30

In conclusion, we have confirmed our assumption that LLMs, acting as posterior transformers, offer learnable and adaptive defenses with good generalization capabilities for different datasets and architectures. They outperform both heuristics and simple noise by providing structured uncertainty and achieving an adequate equilibrium between benefit and privacy.

6.2 Outperforming Previous Defense Approaches

The results of our study show that the three proposed LLM-based posterior encoding approaches substantially boost the robustness of Knowledge Graph Embedding (KGE) models against membership inference attacks on all four KGE algorithms (TransE, TransH, ComplEx, DistMult) and all five evaluated knowledge graph datasets. Specifically, LLM-based Posterior Encoding with Knowledge Distillation achieved the best performance, having the largest reduction in

membership inference attack accuracy. Then, LLM-Guided Posterior Encoding with Secure Aggregation was introduced, and LLM-Guided Posterior Encoding with Noise showed third-best results. In some cases, our methods enabled us to reduce attack accuracy to 0.4, thus outperforming other posterior encoding techniques, such as one-hot encoding, ReLU transformation, shuffling, dropout, top-K selection, variational autoencoders (VAE), and graph perturbation methods, Graph Perturbation 1 (based on Graph Variational Encoder) and Graph Perturbation 2 (based on Graph Autoencoders).

Table 1 presents the minimum values of accuracy in membership inference attacks after employing our security solutions, presented in bold face type.

This outstanding performance can be explained by the possibility of incorporating structured uncertainty in posterior probabilities in a way that would not negatively impact the process of link prediction or entity-relation inference. Unlike other encoding approaches, which are either inefficient at hiding membership signals or generate too much noise to compromise model performance, LLM-based posterior encodings find a compromise between security and prediction performance.

Notably, knowledge distillation outperforms all other techniques, which include secure aggregation and perturbation. Such a trend can be explained by the ability of KD to improve the decision boundaries of the student KGE model, thus preventing overfitting and improving robustness towards membership inference attacks. With the help of soft labels, the student learns the generalized representation of the underlying data, thus minimizing the risk of leakage of the membership-sensitive features. The addition of noise further protects the KGE model, since adversarial models would not be able to draw correct conclusions about membership.

Secure Aggregation performs better than many other techniques; however, in terms of accuracy percentage decrease, it falls just behind KD, primarily because of certain persistent biases of the individual learners that are present in the aggregate forecast. Finally, posterior encoding with only noise provides slightly better results than posterior encoding alone in terms of security; however, the lack of structured knowledge transfer significantly reduces the overall robustness of the approach.

Our posterior encoding techniques based on LLM outperform the traditional techniques as well as graph perturbation methods. This novel approach creates a new framework of protecting the KGE models against membership inference attacks. The results prove the usefulness of using knowledge distillation to enhance robustness of machine learning models.

From the results provided in Fig. 3, one can conclude that the proposed techniques do not reduce the predictive power of the KGE model when applying posteriors; unlike some widely used methods such as Dropout, VAE, GVAE, and GAE whose performance decreases by up to 41.9%. Specifically, the accuracy of Dropout decreases by 30% to 42%, the accuracy of VAE reduces by 7.7% to 13.4%, GVAE (Graph Perturbation 1) shows 14.9% to 23% reduction in per-

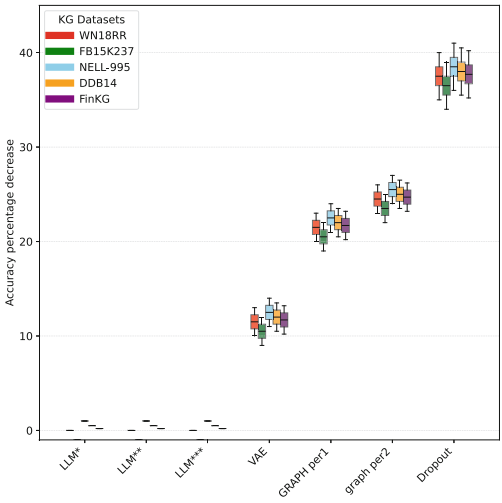


Fig. 3. Accuracy Percentage Decrease in KGs after employing defense approaches

formance, and GAE (Graph Perturbation 2) demonstrates 21% to 29% drop in accuracy percentage.

On the other hand, classical techniques like One-Hot Encoding, ReLU, and Top-K method do not degrade the performance of the KGE model since the maximum probability index is preserved in the process of encoding; hence, there are no problems with class identification based on the triple scoring result. Even though such techniques demonstrate a relatively high level of robustness, they are inferior to our suggested methods, as the latter use advanced approaches such as knowledge distillation and secure aggregation to ensure high classification accuracy and robustness against adversaries' queries.

6.3 Efficiency and Practicality of Proposed Defenses

One of the key benefits of our three proposed defenses that makes them more practical as compared to previous solutions is their efficiency in terms of storage requirements and processing time. Namely, our approaches require only posterior vectors with 3 to 7 elements depending on the number of classes of the considered data set. As opposed to this, methods relying on graph reconstruction such as GAE and VGAE incur heavy costs related to storing and re-calculating the adjacency matrix, which scales in $O(N^2)$. Consequently, we get a transformer with approximately 12 million parameters as opposed to using ensembles and requiring 5 to 10 models or applying GAE/VGAE and doubling the total number of parameters due to duplication of the encoder/decoder.

The training process is made faster as well: our defenses add roughly 7–9% to the training time of KGE models as opposed to baselines. Meanwhile, VGAE training requires 26–30% more processing time due to sampling, GAE

and graph perturbation techniques cause the training time to increase by 39–61%, and using ensembles adds about 5 to 10 times the processing time of KGE training. Although being highly effective at what they do, our LLMs achieve the worst scores for membership inference attack success rates (41–44%) but still remain better than both ineffective but light defenses (e.g., Dropout, Shuffle) and extremely efficient but complex solutions (VGAE, GAE, ensembles). Hence, it can be said that our proposed approaches have the best security-efficiency trade-off when used with posteriors.

7 Conclusion

With the help of our posterior encoding technique based on LLMs, we are able to significantly improve the security of KGE models in terms of protecting them against membership inference attacks while ensuring accuracy remains unchanged. It was revealed that our approaches are more efficient, provide higher resistance, maintain accuracy and improve protection against attacks as opposed to traditional defense solutions (dropout, VAE, GVAE, and GAE). At that, knowledge distillation proved to be especially effective, allowing us to successfully mitigate membership inference attacks without sacrificing the performance of our KGE models.

The future work related to this project will revolve around further refining the proposed solution by experimenting with multiple LLMs other than GPT-4 as well as extending our scope to triple classification and relation prediction tasks. Finally, we aim to study other aspects of the architecture of our model and its posterior probabilities in relation to potential encoding schemes.

References

1. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of EMNLP, pp. 1412–1421 (2015)
2. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
3. Redmon, J., Divvala, S.,Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of CVPR, pp. 779–788 (2016)
4. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. Handbook Brain Theory Neural Netw. **3361**(10), 1995 (1995)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(Dec), 3371–3408 (2010)
7. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Process. Mag. **34**(4), 18–42 (2017)
8. Battaglia, P.W., et al.: Relational inductive biases, deep learning, and graph networks, arXiv preprint arXiv:1806.01261 (2018)

9. Luan, S., Hua, C., Lu, Q., Zhu, J., Chang, X.-W., Precup, D.: When do we need gnn for node classification? arXiv preprint arXiv:2210.16979 (2022)
10. Morshed, M.G. , Sultana, T., Lee, Y.-K.: Lel-gnn: Learnable edge sampling and line based graph neural network for link prediction. IEEE Access (2023)
11. Errica, F., Podda, M., Bacciu, D., Micheli, A.: A fair comparison of graph neural networks for graph classification, arXiv preprint arXiv:1912.09893 (2019)
12. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 1897–1914. IEEE (2022)
13. Wang, Y., Huang, L., Yu, P.S., Sun, L.: Membership inference attacks on knowledge graphs, arXiv preprint arXiv:2104.08273 (2021)
14. Jnaini, A., Koulali, M.-A.: Do LLMs offer a robust defense mechanism against membership inference attacks on graph neural networks? Computers **14**(10), 414 (2025)
15. Olatunji, I.E., Nejdl, W., Khosla, M.: Membership inference attack on graph neural networks. In: 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pp. 11–20. IEEE (2021)
16. He, X., Wen, R., Wu, Y., Backes, M., Shen, Y., Zhang, Y.: Node-level membership inference attacks against graph neural networks, arXiv preprint arXiv:2102.05429 (2021)
17. Jnaini, A., Shirazi, H., Homayouni, H.: Synergy of gpt-3 summarization and vision-encoder-decoder for chest x-ray captioning. In: 2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 476–482. IEEE (2024)
18. Zhong, D., Yu, R., Wu, K., Wang, X., Xu, J., Wang, W.H.: Disparate vulnerability in link inference attacks against graph neural networks. Proc. Priv. Enhancing Technol. **4**, 149–169 (2023)
19. Jnaini, A., Koulali, M.-A.: Can diffusion-based posterior smoothing secure gnns from membership inference attacks? In: 2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA), pp. 1–8. IEEE (2025)
20. Cerda, P., Varoquaux, G., Kégl, B.: Similarity encoding for learning with dirty categorical variables. Mach. Learn. **107**(8–10), 1477–1494 (2018)
21. Wang, Z., et al.: Debiasing learning for membership inference attacks against recommender systems. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1959–1968 (2022)
22. Jnaini, A., Bettar, A., Koulali, M.A. : How powerful are membership inference attacks on graph neural networks? In: Proceedings of the 34th International Conference on Scientific and Statistical Database Management, pp. 1–4 (2022)
23. Singh, A., Thakur, N., Sharma, A.: A review of supervised machine learning algorithms. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1310–1315. IEEE (2016)
24. Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M.: Machine learning-xgboost analysis of language networks to classify patients with epilepsy. Brain Inf. **4**(3), 159–169 (2017)
25. Belgiu, M., Drăguţ, L.: Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote. Sens. **114**, 24–31 (2016)

MobileTestPro: A Real-World Evaluation

Michelle Paredes[1], Danilo Martinez[1,3], Mireya Zapata[2(✉)], and Marcelo Rea[1,3]

[1] Grupo de Investigación en Modelos de Producción de Software (GrIMPSoft), Universidad de Las Fuerzas Armadas ESPE, Av. Rumiñahui, Sangolquí, Ecuador
{meparedes14,mdmartinez,amerea1}@espe.edu.ec

[2] Centro de Investigación en Mecatrónica y Sistemas Interactivos (MIST), Facultad de Ingenierías, Universidad Indoamérica, Quito, Ecuador
mireya.zapata@uti.ec

[3] Instituto Superior Tecnológico Wissen, Av. 10 de Agosto s/n y José Ma. Sánchez, 010107 Cuenca, Ecuador

Abstract. This paper presents the validation of the *MobileTestPro* framework, a systematic, dimension-based approach to assessing the quality of mobile applications. The validation was conducted through an empirical case study involving three application contexts: video streaming, communication, and social media, using YouTube, WhatsApp, and Instagram as units of analysis. The framework assesses key quality attributes, including performance, usability, reliability, and resource consumption, under controlled testing conditions. Specific test cases were designed to simulate real-world scenarios, including variations in video resolution, message exchange processes, and user interaction patterns. Quantitative data were collected and analyzed to compare application behavior across the evaluated dimensions. The results demonstrate that *MobileTestPro* provides a consistent and replicable mechanism for evaluating mobile application quality. Furthermore, the framework facilitates the identification of performance issues and usability limitations. These findings confirm its effectiveness as a structured mobile testing tool.

Keywords: Mobile Testing · Software Quality · MobileTestPro · Case Study · Android · Security Testing · Accessibility · Stress Testing

1 Introduction

The mobile telephony market has experienced exponential growth over the past three decades. Concurrently, the development of mobile applications and platforms has also shown significant evolution, although at a relatively slower pace of expansion. Since the launch of the first iPhone lineup, app development has undergone a complete transformation, making apps more user-friendly and gradually integrating them into everyday life, encompassing all kinds of services: communication, entertainment, and personal and professional productivity. This has brought with it the need to improve app quality in terms of functionality, security, energy consumption, usability, and accessibility.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 38–52, 2027.
https://doi.org/10.1007/978-3-032-30494-0_3

Apps must function in complex environments with a variety of devices, different operating system versions, variations in communication networks, and changes in system policies—what Martinez et al. [1, 2] refer to as "Mobile Illities".

In this context, ensuring the quality of mobile applications is challenging. It is not enough to verify that the application performs the expected actions correctly; it is also necessary to evaluate its behavior under these conditions and test its performance. As an alternative to improve application quality, MobileTestPro [3] emerged. This application evaluation framework consists of three blocks and covers six evaluation dimensions: performance, power consumption, security, accessibility, compatibility, usability, and stress.

MobileTestPro is a framework proposed by Romero-Collaguazo et al. [3] that aims to improve app quality by evaluating six dimensions (energy, usability, security, compatibility, stress, and accessibility). The framework was evaluated using a "toy app", which allowed for verification of its overall functionality and sequence of activities. However, no tests were conducted in real-world environments with widely used applications, where usage conditions are more demanding.

In this article, we present the validation of MobileTestPro in a real-world environment through a case study, applying the guidelines of Runeson et al. [4]. A holistic multicase design was implemented, as proposed by Yin [5]. Three distinct contexts were identified, each with its own unit of analysis (UA). The UAs represent three widely used applications: YouTube (video streaming), WhatsApp (messaging), and Instagram (social media). For data collection, a test case was created for each dimension of MobileTestPro and applied to each UA. The aim was to evaluate whether MobileTestPro allows for the systematic identification of problems and the comparison of results across contexts.

The remainder of the document is organized as follows: Sect. 2 provides background, including the context of MobileTestPro and relevant approaches to mobile testing. While, Sect. 3 describes the case study methodology, its planning, and the research questions. In Sect. 4 we detail the execution of the test case and the results for each unit of analysis. Section 5 discusses the most important findings and their implications. Finally, in Sect. 6, we present conclusions and future work.

2 Background

2.1 MobileTestPro

MobileTestPro is a framework proposed by Romero-Collaguazo [3] that aims to systematize app testing across various quality dimensions, thereby reducing the need for ad-hoc testing and improving the traceability of results. The framework is designed to provide a structure for defining test objectives, activities and evidence to evaluate those attributes which, in the mobile environment, determine the user experience.

MobileTestPro defines a set of activities and tools for each dimension, as shown in Fig. 1, so that testing is carried out optimally in each case and the results can be analysed with greater consistency. This structure facilitates comparison between applications and usage contexts using common criteria, while integrating technical evidence into a framework that supports identifying opportunities for improvement and decision-making during the maintenance and evolution of mobile applications.

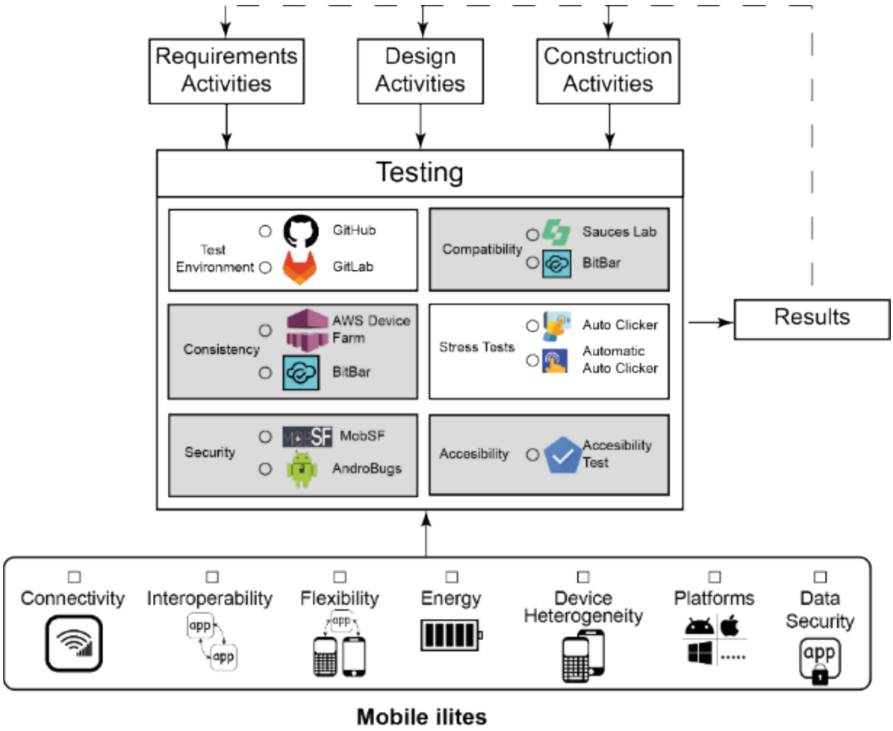


Fig. 1. MobileTestPro (from Romero-Collaguazo): Mobile testing framework to evaluate multiple quality dimensions and mobile-specific attributes.

To test the real effectiveness of MobileTestPro, it is relevant to observe its behavior with apps used daily and by a large number of users, since that is where they have the necessary conditions to obtain data that can confirm its effectiveness in identifying app errors and, therefore, contribute to the best quality of their products.

2.2 Software Testing

According to Myers et al. [6], software testing is a process, or a series of processes, designed to make sure a computer code. Software should be predictable and consistent, offering no surprises to users.

When you want to add some value to a program, you should test it. The reliability of the program means finding and removing errors. The testing process helps to identify errors.

According to Osterweil et al. [7], testing plays an important role in achieving and assessing the quality of a software product. Hence, when we repeat a test-find defects-fix cycle during development, we improve the quality of the product.

Friedman and Voas [8] describe software testing as a verification process for software assessment and improvement. The activities for software quality can be divided into two broad categories: static and dynamic analysis.

Static analysis suggests examining a number of documents, such as requirements documents, software models, design documents, and source code. Traditional static analysis includes code review, inspection, walkthrough, algorithm analysis and proof of correctness.

Dynamic analysis of a software application includes its execution to expose possible program failures. Furthermore, its behavioural and performance are also observed. Programs are executed with a set of input values and the outputs are observed. In testing process we observe some representative behaviors and reach a conclusion about the its quality.

A test case is a simply pair data, input data and outcome expected.

Testing Activities

In order to test a program, we need a set of sequence activities. Naik et al. [8] proposed the following steps.

- Identify an objective to be tested: Identify the objective to be tested. Define the purpose of the designed one or most test cases to ensure that the program supports the objective.
- Select inputs: Selection of test inputs based on the requirements or goal of the application.
- Compute the expected outcomes: Calculate or define the outcomes of the applications with the selected inputs.
- Set up the execution environment: Prepare the scenario to execute the program that includes setting up the run environment (hardware, software, and communication).
- Execute program: Run the programs with selected inputs and observe the outcomes of the programs.
- Analyze results: The final activity is to observe the outcomes and compare them with the expected results.

2.3 Mobile Testing

Mobile apps, like any other software system, require testing, but apps are different from traditional ones and require different and specialized testing techniques and processes [9]. The difference between traditional applications and mobile applications Martinez [1] et al. named Mobile Ilities.

Haller [10] presents device testing and real-world testing as a dual approach for testing applications. Device test consist in testing the same test case (app) in different mobile devices to ensure compatibility, performance, usability, etc. While testing in the wild approach is testing an app in its real usage context, on other hands, in a real environmental.

When we used Mobile Ilities and Haller's approach, we obtained the best conditions for applying MobileTestPro in our study, because we created a set of test cases based on Mobile Ilities to apply to MobileTestPro's dimensions. In the test, we apply two Haller approaches. Any case test was testing on some devices (physical or emulator) on a real-world application in its real environment through a case study.

3 Case Study

3.1 Case Study Design

To study the behavior of MobileTestPro in a real environment, a case study was proposed using the guidelines proposed by Runeson et al. [4] because it is the most appropriate method to study the behavior of a phenomenon in its real context, especially when the conditions of the environment directly influence the observed behavior, and it is not possible to control all variables.

The following research questions were posed for the study:

RQ1: How effectively can MobileTestPro identify performance, stability, security, usability, compatibility, and accessibility issues in applications?
RQ2: How do the dimensions of MobileTestPro relate to the identification of recurring failures and the comprehensive evaluation of mobile application quality?

For the study, the mobile applications that register the highest use according to the DataReportal [11] have been considered, being Instagram, YouTube, and WhatsApp, which in turn have determined that the study is carried out in 3 contexts (streaming, communication, and social networks).

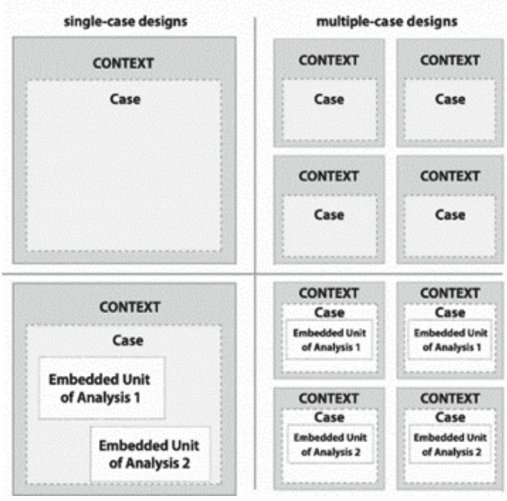


Fig. 2. Use Case Design (from Robert Yin): Types of case study designs proposed by Yim [5] according to the defined contexts and units of analysis.

The design adopted is a holistic, multi-case approach, as described by Yim et al. [5] as shown in Fig. 2. It is multi-case because three contexts are analyzed (video Streaming, Communication, and social media), and it is holistic because each context has its unit of analysis, see Fig. 3. This design allows for comparing the performance of MobileTestPro across different functional domains and observing how dimensions such as usability, compatibility, accessibility, security, performance, and stability interact in everyday scenarios.

Fig. 3. Holistic Multiple-case Design: Design composed of 3 contexts (video streaming, communication, social media), with a unit of analysis for each context (YouTube, WhatsApp, Instagram).

3.2 Data Collection

For data collection, a set of test cases was developed for each of the units of analysis. The test cases were planned based on the context and the MobileTestPro dimensions that best suited the nature of the application.

The planned test cases for the YouTube analytics unit are shown in Table 1.

Table 1. YouTube planned test cases

Dimension	Objective	Input	Outcome	Tool
Compatibility and Usability	Verify that YouTube is compatible and usable on different device models and Android versions	The YouTube APK runs on available devices and emulators	The app should be installed, open, and allow basic navigation without errors on the selected devices	BitBar
Performance and energy	Evaluate YouTube's behavior during continuous video playback in different qualities	5 and 10-minute video playback sessions in different qualities	The device must maintain stable operation, without unexpected shutdowns or appreciable performance degradation	AccuBattery
Stress	Simultaneous user loading on services exposed by the YouTube API	K6 scripts configured with different levels of concurrent users	The invoked services must maintain response times within acceptable parameters for the defined test scenario	k6

(*continued*)

Table 1. (*continued*)

Dimension	Objective	Input	Outcome	Tool
Security	Static analysis of the YouTube APK to identify potential security risks	YouTube APK loaded into the tool	Detect vulnerabilities classified as critical in the static analysis report	MobSF
Accessibility	Check compliance with basic accessibility criteria on the main YouTube screens	Home, search, and playback screens	Interactive elements and texts must comply with basic guidelines for contrast, size, and labeling	Accessibility Test/Accessibility Scanner

Meanwhile, for the WhatsApp analysis unit, the test cases shown in the Table 2.

Table 2. WhatsApp planned test cases

Dimension	Input	Outcome	Objective	Tool
Compatibility and Usability	Check WhatsApp compatibility and the correct display of the messaging interface on different Android devices	WhatsApp APK running on multiple devices	App must correctly display the main screens and allow access to chats without interface failures	BitBar
Performance and energy	Analyze WhatsApp's resource consumption during active messaging use	Sessions of 5 and 10 minutes, sending text messages, voice notes, and images	App should respond smoothly and the operating system should not exhibit any obvious crashes or delays	AccuBattery
Stress	Evaluate the API's behavior when the app is sending and querying messages under load	k6 scripts executing concurrent calls to selected endpoints	The system must support the defined load without exhibiting widespread failures in its responses	k6

(*continued*)

Table 2. (*continued*)

Dimension	Input	Outcome	Objective	Tool
Security	Review permissions, settings, and potential security weaknesses in the WhatsApp APK	WhatsApp APK analyzed in the tool	The report should not reveal clearly insecure configurations or practices at the level of permissions and data storage	MobSF
Accessibility	Evaluate the accessibility of WhatsApp's chat interface and main menus	Chat list screen and conversation view under analysis	Buttons, fields and interactive elements must be identifiable and accessible via the tool	Accessibility Test / Accessibility Scanner

Finally, for the Instagram analysis unit, the test cases shown in the Table 3.

Table 3. Instagram planned tests

Dimensión	Objective	Input	Outcome	Tool
Compatibility and Usability	Check Instagram's compatibility with dynamic layouts across different screen sizes and Android versions	The Instagram app runs on devices with different screen sizes and resolutions and versions of Android	The feed and main screens must load correctly and respond to basic user interactions	BitBar
Performance and energy	Evaluate Instagram's performance during continuous Reels viewing	5 and 10-minute Reels viewing sessions	Navigation should remain smooth, and the application should not crash or freeze during continuous use	AccuBattery
Stress	Validate the stability of the API associated with the Instagram feed and authentication under load scenarios	K6 scripts with concurrent users accessing content query endpoints	Services must maintain consistent availability and response times throughout the test execution	k6

(*continued*)

Table 3. (*continued*)

Dimensión	Objective	Input	Outcome	Tool
Security	Evaluate the Instagram APK to identify potential vulnerabilities in internal components	Instagram APK processed in the tool	High-impact vulnerabilities that compromise data integrity or confidentiality should not be registered	MobSF
Accessibility	Check accessibility in the feed and dynamic content sections, including Reels.	Feed and Reels screens analyzed with the accessibility tool	The interface components must be navigable, readable, and have appropriate labels for assistive technologies	Accessibility Test / Accessibility Scanner

3.3 Execution

For the tests, a series of free simulators was used, as well as a smartphone running Android 12, a Unisoc Tiger T606 processor, a 6.5-inch screen, 4 GB of RAM, and a 500 mAh battery. The phone was prepared exclusively for testing; therefore, factory settings were applied, and only the applications necessary for testing were installed.

The test cases were executed by applying the MobileTestPro dimensions for each usage context and selected unit of analysis. To ensure consistency and comparability between cases, each application was evaluated following the same execution logic: (i) definition of the use scenario corresponding to the context, (ii) execution of test cases by dimension using specific tools, and (iii) systematic recording of evidence for later consolidation.

The planned test cases were executed by dimensions, each associated with a mechanism for obtaining evidence: compatibility using BitBar; security using MobSF; performance under load using k6; energy consumption using AccuBattery; and accessibility using an accessibility scanner.

4 Results

The results obtained from running the planned test case are shown in this section.

4.1 Video Streaming

The planned test cases designed for this analysis unit was carried out in two parts. For the first part, a 5-minute video with a resolution of 1080 was used. For the second part,

another video with a higher resolution and a duration of 10 min was used. The results obtained for each dimension are shown in the Fig. 4.

Fig. 4. UA YouTube: Results obtained after performing the YouTube UA test cases in the planned dimensions.

Table 4 shows the details of the test cases execution for each dimension of MobileTestPro.

Table 4. Details of tests performed on YouTube

Dimension	Details
Compatibility (BitBar)	YouTube worked correctly on the Pixel 3a (Android 12). On the Pixel 9 Pro (Android 14), it was reported to be incompatible, and the error INSTALL_FAILED_MISSING_SPLIT was recorded
Security (MobSF):	The analysis reported a score of 52/100 (classification: medium risk-go) and 4 vulnerabilities: Cleartext Traffic, use of MD5, exported components and enableBackup = true
Stress (k6):	The test showed approximately 81% successful results and 19% failures, with connection termination being the most common failure
Energy (AccuBattery):	In the YT-02 scenario (720p, 10 min), an average power consumption of 592 mA was recorded
Accessibility	A total of 49 suggestions were identified, the most notable of which concerned opportunities for improvement in touchpoints and in the text or labels

4.2 Communications

The test cases carried out by WhatsApp's analytics team were conducted on two devices running different versions of the Android operating system. The test involved sending text messages, photos, videos, and voice notes in rapid succession over two time periods: the first lasting 5 min and the second 10 min. The results are shown in Fig. 5. Table 5 shows the details of the test results for this unit of analysis.

Fig. 5. UA: WhatsApp: Results obtained after performing the WhatsApp UA test cases in the planned dimensions.

Table 5. Details of tests performed on WhatsApp

Dimension	Details
Compatibility (BitBar)	WhatsApp was reported to be compatible on the Pixel 3a (Android 12) and the Nexus 5 (Android 7.1.1). In the latter case, CPU usage was recorded at approximately 80–150%
Security (MobSF):	The analysis yielded a score of 51/100 (classification: medium risk). The findings reported include: unencrypted traffic, user certificates, SHA1 + V1 and hard-coded secrets
Stress (k6):	The results showed approximately 37% correct answers and 62% errors, with connection errors being the most common
Energy (AccuBattery):	In scenario WA02 (10 min), an average current consumption of 691.2 mA was recorded
Accessibility	A total of 35 suggestions were recorded, with frequent improvements to touch targets and text/labels

4.3 Social Media

Finally, for Instagram, test cases were designed to assess compatibility across different devices and versions of Android. In addition, installation and execution were verified, and CPU and memory usage was monitored during startup on the BitBar platform. The results are shown in Fig. 6.

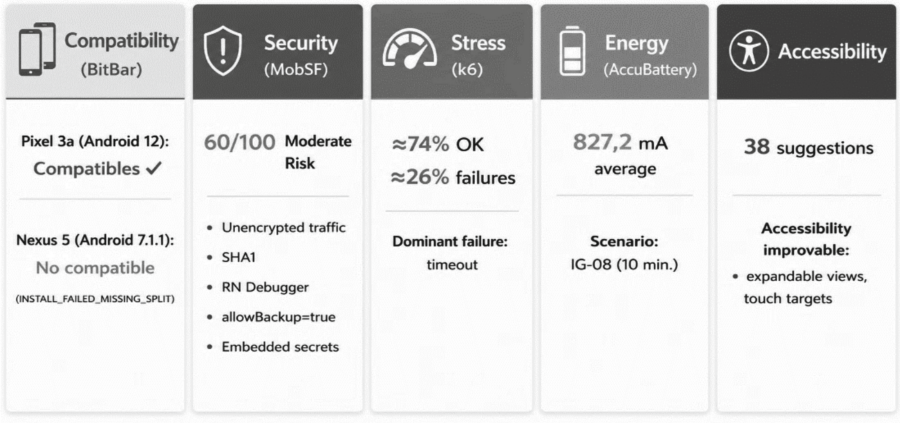


Fig. 6. UA: Instagram: Results obtained after performing the Instagram UA test cases in the planned dimensions.

Table 6 shows the details of the results obtained during the testing of this analysis unit.

Table 6. Details of tests performed on Instagram

Dimension	Details
Compatibility (BitBar)	It has been reported that Instagram is compatible with the Pixel 3a (Android 12). On the Nexus 5 (Android 7.1.1), it has been reported that it is not compatible, and the error INSTALL_FAILED_MISSING_SPLIT has been recorded.
Security (MobSF):	The analysis returned a score of 60/100 (classification: medium risk), with findings related to: unencrypted traffic, SHA1, the presence of RN Debugger, and the setting allowBackup = true
Stress (k6):	The results showed approximately 74% correct answers and 26% errors, with delay time being the most common error
Energy (AccuBattery):	In scenario IG-08 (10 min), an average current consumption of 827.2 mA was recorded
Accessibility	A total of 38 suggestions were received, with expandable views and touch targets standing out as areas for improvement

5 Discussion

The data obtained during the case study have enabled us to identify several faults in the performance of the evaluated apps, demonstrating the effectiveness of the framework. The findings suggest that the dimensions do not function as isolated assessments; rather, they tend to converge into consistent relationships that facilitate the interpretation of the

system's quality. Furthermore, a number of recurring faults were identified among the apps evaluated.

Security is a cross-cutting issue linked to the robustness and reliability of the system. In this area, it was found that the "allowBackup = true" vulnerability is present on both YouTube and Instagram. These apps require users to register an account in order to run. This vulnerability allows backup data to be stored on the device, which can then be stolen from it via some form of attack. Consequently, users are exposed to identity theft, data exposure, social engineering, scams and extortion, amongst other risks.

Furthermore, unencrypted traffic affects both WhatsApp and Instagram, as both apps use encryption protocols such as HTTPS/TLS to send messages, photos, videos, credentials and session tokens. The vulnerability can arise when apps fail to validate certificates correctly, use insecure TLS configurations, do not implement 'certificate pinning', or allow 'man-in-the-middle' attacks. The risks associated with incorrect encryption implementation include interception of traffic between the user and the server via an attack, which could lead to the disclosure of transmitted information, modification of requests, theft of authentication tokens, and session hijacking.

The results show a close relationship between compatibility and accessibility. Both aspects are reflected in user interaction with the interface and often manifest as similar issues when the environment changes (such as changes in resolution, pixel density or operating system version). In this regard, issues such as small touch targets or insufficient labels not only affect users with accessibility needs but also tend to become more critical when the application runs on different configurations. Therefore, the evidence supports the view that compatibility and accessibility share a pattern: when the interface is not optimised, the impact is amplified across multiple runtime environments.

In terms of efficiency and resource usage, a correlation was observed between stress and energy consumption, as both factors reflect the system's behaviour when resource demand increases. Under higher-load scenarios, increased activity (CPU and network) can result in a greater energy impact; therefore, a combined assessment makes it possible to determine not only whether the application responds, but also whether it does so whilst making efficient use of the device's resources.

6 Threats to Validity

The study operationalizes mobile application quality using the dimensions defined by MobileTestPro; however, the selected test cases and tools may not fully capture the complexity of real-world usage. In particular, stress testing was based on simulated API calls rather than actual user behavior, which may limit the realism of the observed system responses. The execution protocol was consistent, the use of predefined scenarios may bias defect detection compared to more flexible approaches such as ad-hoc or exploratory testing. The absence of a direct comparison with a baseline approach (e.g., measuring time to detect defects using MobileTestPro versus ad-hoc testing) also restricts the ability to attribute efficiency gains solely to the framework.

The generalizability of the findings is constrained by the limited number of devices (2–3 devices and emulators), which does not fully represent the diversity of the mobile ecosystem, including variations in screen sizes, aspect ratios, hardware configurations,

and emerging form factors such as foldable devices. Similarly, the evaluation is restricted to three applications and contexts, which may not capture the full spectrum of mobile application behaviors. In terms of robustness, the results are based on a limited set of executions without extensive variability in conditions such as network performance or repeated trials. Expanding device coverage, incorporating real user load scenarios, and including comparative and statistical analyses would improve the validity and reliability of the findings.

7 Conclusion and Future Works

MobileTestPro was validated through a holistic multi-case study applied to three usage contexts (video streaming, communications and social media), represented by the following analysis units YouTube, WhatsApp and Instagram. The results show that the dimensional approach enabled tests to be planned and executed systematically, evidence to be collected that was comparable across units of analysis, and the detection of recurring patterns relevant to the evaluation of mobile applications in real-world scenarios.

The findings show that these dimensions cannot be interpreted in isolation; compatibility and accessibility tend to reflect common interface and interaction issues. Furthermore, stress and energy consumption provide complementary evidence on resource usage under greater demand. Finally, the security dimension enables the identification of risks associated with configuration and best practices, even when these do not manifest as visible failures to the user.

Taken together, these results confirm that MobileTestPro is useful for organising and comparing evaluation by dimension in scenarios that closely resemble real-world use.

Future work includes expanding the validation coverage with more execution configurations (more models and system versions) and increasing the number of applications based on their context. Furthermore, it is recommended to strengthen the synthesis of results across dimensions using integration criteria that allow for more direct case comparisons and support improvement decisions.

References

1. Martinez, D., Ferre, X., Guerrero, G., Juristo, N.: An agile-based integrated framework for mobile application development considering Ilities. IEEE Access. **8**, 72461–72470 (2020). https://doi.org/10.1109/access.2020.2987882
2. Martínez, D., Ferré, X., Marcillo, D.: Proposal for an integrated framework for mobile applications development. In: 2018 International Conference on Information Technology & Systems, Santa Elena, pp. 813–819 (2018). https://doi.org/10.1007/978-3-319-73450-7_77
3. Romero-Collaguazo, J., Martinez, D., Zapata, M., Ferre, X.: MobileTestPro: testing framework for mobile application. In: Gervasi, O., et al. (eds.) ICCSA 2025. LNCS, vol. 15650, pp. 68–80. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-96962-1_5
4. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empir. Softw. Eng. **14**, 131–164 (2009). https://doi.org/10.1007/s10664-008-9102-8
5. Yin, R.K.: Case Study Research. Design and Methods, vol. 4, pp. 264–267. Sage (2009). https://doi.org/10.1007/BF01103312

6. Myers, G.J., Badgett, Thomas, T.M., Sandler, C.: The Art of Software Testing. Wiley, Hoboken (2004)
7. Osterweil, L.: Strategic directions in software quality. ACM Comput. Surv. **28**, 738–750 (1996). https://doi.org/10.1145/242223.242288
8. Naik, K., Tripathy, P.: Software Testing and Quality Assurance: Theory and Practice
9. Muccini, H., Di Francesco, A., Esposito, P.: Software testing of mobile applications: challenges and future research directions. IEEE (2012)
10. Haller, K.: Mobile testing. ACM SIGSOFT Softw. Eng. Notes **38**, 1–8 (2013). https://doi.org/10.1145/2532780.2532813
11. Global Social Media Statistics — DataReportal – Global Digital Insights. https://datareportal.com/social-media-users. Accessed 12 Mar 2026

ABCD-Based Melanoma Detection in Non-dermoscopic RGB Imaging

Maria Oniga(✉), Paul Florin Rus, and Alina Elena Sultana

National University of Science and Technology POLITEHNICA Bucharest, Bucharest 060042, Romania
{maria.oniga,alina_elena.sultana}@upb.ro, paul_florin.rus@stud.acs.upb.ro

Abstract. Melanoma is a highly aggressive skin cancer whose prognosis strongly depends on early detection. The ABCD rule of dermatology provides a clinically interpretable framework for lesion assessment based on asymmetry, border irregularity, color variation, and diameter. However, most existing computational implementations rely on dermoscopic images acquired under controlled conditions, limiting their applicability in real-world settings.

In this work, we propose a computational framework for the extraction of ABCD features from non-dermoscopic images captured using a standard RGB camera at a fixed distance. The proposed method includes an illumination correction pipeline, localization of the lesion, and mask creation. In addition, the study includes a color model derived from dermoscopic images based on color quantization and hybrid color classification, which adapts to the used dataset. The asymmetry and border irregularity scores were computed using geometric descriptors, while the diameter was determined as a feature and converted into mm.

The four features obtained were further used to create an adapted Total Dermoscopic Score (TDS), which classifies the skin lesions as benign, dysplastic, or malignant. The proposed approach was evaluated on a proprietary dataset of 146 images obtained using a standard camera. The overall accuracy obtained was 70.55%, which demonstrates that the ABCD rule can be adapted to standard RGB imaging, providing a low-cost alternative for melanocytic lesions in non-dermoscopic conditions.

1 Introduction

Melanoma is an aggressive form of skin cancer that arises from abnormal proliferating pigment-producing cells named melanocytes. [11]. Currently, this skin disease is a public health concern, whose incidence and mortality are increasing at an alarming rate [10]. According to the World Health Organization, in 2022, melanoma was diagnosed in 331,722 new cases and was responsible for 58,667 deaths globally, with a 2025 projection of 1,092,818 new cases in 2025 [4]. Thus, early detection is essential to reduce mortality rates.

Melanocytic lesions are diagnosed through dermoscopy, an examination using adequate illumination and magnification [14]. This technique is an essential tool

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 53–70, 2027.
https://doi.org/10.1007/978-3-032-30494-0_4

for lesion differentiation, which enhances the visualization of superficial and subsurface structures and improves diagnostic accuracy. Dermoscopes use polarized light to reduce reflection and enhance contrast, facilitating the visualization of diagnostically relevant features [5].

From a clinical perspective, melanoma is a pigmented lesion that is characterized by changes in size, shape, and color. These features are determined by the ABCDE criteria (Asymmetry, Border irregularity, Color variation, Diameter, and Dermoscopic features and Evolution), which guide the initial lesion assessment [3]; however, histopathological examination remains the preferred method to determine the final diagnosis [14].

The four criteria from the ABCD rule are used to determine the Total Dermoscopic Score (TDS), which provides a standardized pipeline for classifying skin lesions as benign, suspicious, or malignant. Thus, it is employed frequently in computer-aided diagnosis systems, where it enables quantitative analysis of specific features and improves diagnostic accuracy [2].

Recent studies on automated melanoma detection can be broadly categorized based on the type of imaging modality used and the level of interpretability of the extracted features.

Several studies rely on dermoscopic images, where controlled acquisition conditions and magnification facilitate the extraction of clinically meaningful structures. In this setting, several classical image-processing approaches quantify the ABCD rule. Study [16] computes asymmetry via vertical/horizontal partitioning with entropy and bi-fold measures, detects border irregularity using Sobel edge analysis, quantifies color with six RGB triplets and Euclidean thresholds, and converts diameter from pixels to millimeters. Dermoscopic images were also used in [12] to extract asymmetry through diagonal partitioning, evaluate border irregularity across eight regions, combine color occurrence with RGB statistics, and interpret diameter relative to the 6 mm threshold. Likewise, [7] applies the ABCD rule using shape/color asymmetry, eight-region border analysis, and RGB-based quantification of six color hues. More recent dermoscopy-based studies preserve the same interpretability while integrating learning-based classification stages. For example, paper [8] combines lesion segmentation, explicit ABCD feature extraction, and artificial neural networks for skin tumor classification using dermoscopic images. Lesion evolution can be quantified by using sequential dermoscopic analysis by capturing the changes in morphology and color over time [15]. Recent deep learning methodologies attempted to maintain interpretability by also predicting lesion classification and clinically significant ABCDE criteria from dermoscopic images [6].

On the other hand, approaches based on non-dermoscopic images are significantly limited and encounter more limitations. The study [1] proposed a two-stage methodology for standard camera images, adjusting the ABCD rule. Rather than using the classical dermoscopic formulation, their framework implemented a large set of geometric, gradient-based, and statistical color descriptors.

The ABCD rule is commonly used in dermoscopic imaging, but its direct application to standard RGB images remains problematic due to inconsistencies

in acquisition conditions, which can alter color consistency and feature reliability. In this context, we propose a clinically inspired approximation of ABCD features for standard RGB images captured in real-world clinical settings, thus improving dermatological interpretability beyond dermoscopy. The main question is whether ABCD-derived features can maintain diagnostically significant patterns in the absence of dermoscopic structures. Instead of directly copying dermoscopic criteria, we change them so that they still mean the same thing when it comes to asymmetry, border irregularity, color variation, and diameter, while also taking into account the differences that come with standard RGB imaging.

An important aspect of this work was the development of a feature extraction pipeline that can handle changes in lighting and color, which are difficult to avoid in diverse clinical settings. Non-dermoscopic imaging has a lot of differences in lighting, color balance, and reflectance, which differ significantly from dermoscopic systems, where acquisition conditions are controlled. To address this, we propose a preprocessing and color modeling method that is illumination agnostic by using gray-world normalization, reference-based white balance, and color calibration.

The proposed method ensures that the ABCD criteria can be adapted to standard RGB imaging by modeling each feature accordingly. A reference marker ensures the computation of the diameter in millimeters, while the measures based on contour and symmetry capture lesion morphology. The color variation is computed using a hybrid approach that can determine color patterns in non-dermoscopic images by using dermoscopic knowledge.

Overall, the proposed method allows the development of low-cost, easy-to-understand tools for melanoma screening by showing that ABCD-derived features can still show meaningful diagnostic trends without dermoscopic detail. This is especially important for use in a variety of clinical settings where standardized imaging is not always available and the ability to handle acquisition variability is crucial.

2 Methodology

This study proposes a framework for the assessment of melanocytic lesions based on the ABCD rule of dermatology using standard RGB images. Unlike conventional approaches relying on dermoscopic imaging, the proposed method operates on images acquired with a standard camera at an approximate distance of 40 cm. The framework extracts and quantifies the four components of the ABCD rule, computes the TDS, and enables the identification of the blue veil for lesion classification (see Fig. 1).

2.1 Data Acquisition

The images used in this study were acquired using the MelanoDet system, first presented in [13]. MelanoDet (Fig. 2) is a multispectral imaging platform com-

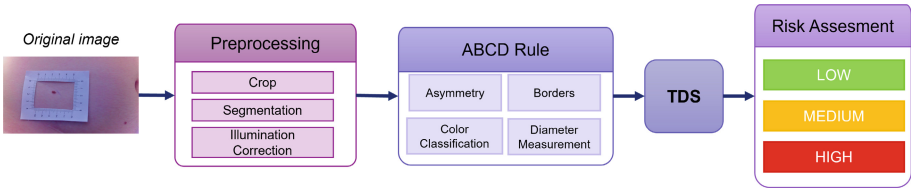


Fig. 1. Schematic diagram of the proposed method.

prising three cameras operating in different spectral ranges: visible (VIS), near-infrared (NIR), and long-wavelength infrared (LWIR, thermal imaging).

In this work, only the VIS modality was considered. A key advantage of the VIS camera within this system is its ability to support accurate lesion localization and to provide detailed morphological information, including color distribution, shape, and lesion area. These features enable the extraction of clinically relevant descriptors that contribute to the differentiation between benign and dysplastic lesions.

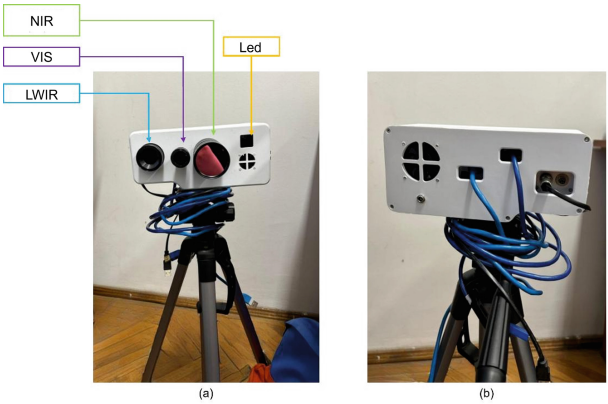


Fig. 2. MelanoDet imaging system.

Since this study focuses on non-dermoscopic RGB images, only the VIS camera specifications are considered. The camera module has a resolution of 1920×1080 pixels and is based on a 1/2" Sony IMX385 CMOS sensor, offering high sensitivity and high dynamic range (HDR). It provides a wide field of view (DFOV: 122°, HFOV: 110°) and supports object detection at working distances from approximately 30 cm to infinity. The module is equipped with an RGB M12-mount varifocal lens that allows manual focus adjustment.

2.2 Dataset

The dataset used in this study was acquired using a multispectral device presented in [13], over a period of 11 months, in collaboration with the Central

Military Emergency University Hospital "Dr. Carol Davila", Bucharest, Romania, and St. John's University Hospital, Bucharest, Romania. All participants provided informed consent, and the study was conducted in accordance with the Declaration of Helsinki (1975, revised 2004) and approved by the National University of Science and Technology POLITEHNICA Bucharest, University Ethics Committee, Subcommittee for Research Ethics (approval no. 26/06.10.2025).

The images were acquired from 40 cm, which ensured a consistent field of view across patients, without the need of magnification. In addition, for each patient a reference marker was used to enclose the lesion with the final purpose of localizing the lesion and perform the conversion from pixels to mm. Since the image acquisition was performed in a clinical setting, the illumination conditions were not controlled. This resulted in variability across images in terms of lighting, but consistent in positioning and framing of the lesion.

A total of 146 cases were collected, with diagnoses established either by experienced dermatologists or through histopathological analysis. For each case, three images were acquired in the VIS, NIR, and LWIR spectra, resulting in a total of 438 images. The dataset is categorized into three lesion classes: nevus (NV), dysplastic nevus (DN), and melanoma (MM). The sample distribution for each class is depicted in Fig. 3. In line with the objective of this study, which focuses on applying the ABCD rule to non-dermoscopic imaging, only VIS images were selected for further analysis.

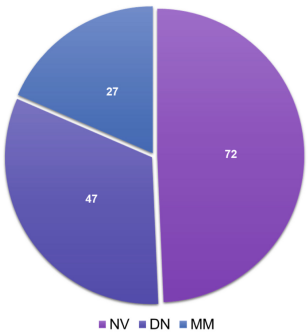


Fig. 3. Class distribution within the dataset.

2.3 Preprocessing

Since the images were acquired using an RGB camera in a clinical environment, where strict illumination control is difficult to enforce, a preprocessing pipeline was introduced to reduce variability in lighting and ensure color consistency across the dataset.

Lesion localization was performed via an initial cropping step using a pretrained YOLOv8n object detection model. Initialized with COCO weights and

fine-tuned on 50 manually annotated images to detect the reference marker, the model was trained on the VIS modality with data augmentation for 20 epochs. During inference, predictions were filtered with a 0.5 confidence threshold and 0.3 IoU-based non-maximum suppression; the highest-confidence bounding box was selected, expanded by a factor of 1.2, and used to extract the Region of Interest (ROI) for subsequent processing.

An illumination correction step was applied, combining gray-world color constancy with white balance adjustment. The gray-world assumption normalized the global color distribution to reduce color cast, after which white balance was performed relative to the red channel to preserve the natural color of human skin. The reference marker in each image was used to identify a white reference region (bright areas were detected via intensity thresholding, and candidate patches were extracted from the marker). From the patches extracted, the most neutral one (where there was minimal variation in the R,G,B channels mean) was selected as a reference for the white balance processing. Since the skin is characterized by red tones, we further computed the scaling factors for the green and blue channels relative to the red channel, and applied them to the entire image.

Since most of the images in the dataset contain small or low-contrasted lesions, automated segmentation methods were not reliable. Hence, we manually segmented each image and created masks using Adobe Photoshop. To ensure that the final mask contained only the lesion, we further applied a morphological erosion with a 5×5 structuring element.

2.4 ABCD Rule

The ABCD rule of dermatology is a widely used clinical guideline for the assessment of melanocytic skin lesions and early detection of melanoma. It is based on four criteria, each reflecting structural characteristics associated with malignancy.

Asymmetry (A) Asymmetry was determined along the vertical and horizontal axes, using a crop of the lesion mask which was further divided in two halves, corresponding to each axes.

Vertical asymmetry was computed by splitting the lesion into the left half (L) and the right half (R). The right half was flipped and compared to the left half by measuring how much they overlapped:

$$A_v = 1 - \frac{|L \cap R|}{|L \cup R|} \tag{1}$$

A similar process was performed for the horizontal asymmetry, where the lesion was divided into a top half (T) and a bottom half (B):

$$A_h = 1 - \frac{|T \cap B|}{|T \cup B|} \tag{2}$$

The overall asymmetry measure (A_{mean}) was defined as:

$$A_{mean} = \frac{A_v + A_h}{2} \tag{3}$$

We used a search process to find two thresholds (t_1, t_2) that would allow the mapping of the overall asymmetry values to the ABCD scoring system. In particular, all possible thresholds were evenly sampled in the $[0, 1]$ range, and all valid pairs (t_1, t_2), $t_2 > t_1$ were verified. We calculated the asymmetry score for all lesions in each pair and then compared the ABCD-based classification to the ground-truth labels. The best thresholds were chosen as those that gave the dataset the highest overall classification accuracy. Thus, the final values were $t_1 = 0.4$ and $t_2 = 0.7$.

Using the determined thresholds and considering that the asymmetry score ranges between 0 and 2, the final score A was defined as 0 if $A_{mean} < t_1$, 1 if $t_1 \leq A_{mean} < t_2$, and 2 if $A_{mean} \geq t_2$ [16].

Border Irregularity (B). Border irregularity was quantified using two complementary shape descriptors computed from the binary lesion mask: convexity irregularity and radial distance variation.

First, the external contour of the lesion was extracted, and the largest contour was retained for analysis. To measure global deviation from a regular shape, the convex hull of the lesion contour was computed. The convexity irregularity was then defined as

$$I_c = \frac{A_{\text{hull}} - A_{\text{lesion}}}{A_{\text{hull}}}, \tag{4}$$

where A_{lesion} is the contour area of the lesion and A_{hull} is the area of its convex hull.

To assess variations in the local contour, the centroid of the lesion was calculated using contour moments, and the Euclidean distance from the centroid to each contour point was evaluated. The coefficient of variation of the radial distances was employed to characterize the radial irregularity:

$$I_r = \frac{\sigma(d)}{\mu(d)}, \tag{5}$$

where $\sigma(d)$ and $\mu(d)$ represent the standard deviation and mean of the set of radial distances.

The final border irregularity measure was obtained by equally combining I_c and I_r:

$$I_b = 0.5 \cdot I_c + 0.5 \cdot I_r. \tag{6}$$

Finally, linear rescaling was used to map the continuous irregularity score to the ABCD border scale. The resulting value was then clipped to the interval $[0, 7]$, which lets it match the standard ABCD scoring range for border irregularity and be directly integrated into TDS [16].

Color (C). The first color analysis used six clinically important dermoscopic colors: white, red, light brown, dark brown, blue-gray, and black. These colors are often used in dermatological evaluations and are described in the literature with reference RGB triplets and their corresponding thresholds.

To perform color classification, we used a reference RGB triplet along with a RGB threshold for each of the six colors analyzed. If the Euclidean distance between a pixel's normalized RGB value and the reference color was less than the threshold, the pixel was assigned to that color [16].

Table 1 shows a summary of the reference values and thresholds that were used in the first implementation.

Table 1. RGB triplets used as reference and thresholds for the six known dermoscopic colors [16]

Color	Reference	Threshold
White	(1, 1, 1)	(0.8039, 0.8039, 0.8039)
Red	(0.8, 0.2, 0.2)	(0.5882, 0.1961, 0.1961)
Light Brown	(0.6, 0.4, 0.25)	(0.5882, 0.1961, 0)
Dark Brown	(0.4, 0.26, 0.13)	(0.1961, 0, 0)
Blue Gray	(0.2, 0.6, 1)	(0, 0.3922, 0.5882)
Black	(0, 0, 0)	(0.1961, 0.1961, 0.1961)

Practically, for each pixel p and color reference c_k, the distance was computed as:

$$d_k = \|p - c_k\|_2 \tag{7}$$

A pixel was assigned to color k if:

$$d_k \leq T_k \tag{8}$$

where T_k represents the threshold associated with the color k [16].

These reference values are established for dermoscopic images obtained under controlled conditions; however, this study utilized standard RGB images captured at 40 cm under clinical conditions, resulting in variability in illumination, color balance, and sensor response. As a result, direct application caused problems, such as misclassifying red and brown tones and not being able to find blue-gray areas very well. This led to a data-driven color refinement strategy.

Hybrid Color Classification To adapt dermoscopic color definitions to non-dermoscopic images, a data-driven calibration was performed using color quantization, where K-means clustering was applied to lesion pixels to extract dominant colors and refine RGB triplets. Color assignment thresholds were then empirically determined from pixel-reference distance distributions, balancing sensitivity and specificity while minimizing noise.

A hybrid classification strategy was employed, in which an initial rule-based stage defined color categories using RGB channel relationships, with red characterized by a dominant R channel and sufficient saturation while excluding brown-like patterns, brown tones by $R > G > B$, blue-gray by $B > R$ and $B > G$, and black or white determined using intensity-based thresholds. Pixels not classified in this stage were assigned using Euclidean distance to reference colors.

To reduce the effect of noisy pixels on color identification, a relative threshold based on the lesion area was added. Colors that made up less than a certain percentage of the lesion were left out because they were very few and still counted as separate colors. We looked at a few different thresholds (0%, 0.3%, 0.5%, and 0.8%) to see how they affected the final TDS and the performance of the classification.

Finally, the color score C was defined as the number of distinct colors present after applying the proposed classification algorithm, with the C score ranging from 1–6 [16].

Color Uniformity. We also determined the color uniformity to see how the colors were spread out in the lesion based on the number of pixels assigned to each color. If a dominant color covered more than 75% of the lesion area or more than 60% of the area with the second most common color below 30%, the lesion was considered uniform. This measure shows how similar the colors are. Most of the time, benign lesions have colors that are more even, while melanoma has colors that are more varied. Even though it wasn't part of the traditional ABCD rule, this metric was used to help figure out how colors changed.

Diameter. The lesion diameter was estimated using the equivalent diameter derived from the lesion area:

$$D_{px} = \sqrt{\frac{4A}{\pi}} \tag{9}$$

The diameter was converted from pixels to millimeters using a calibration factor:

$$D_{mm} = \frac{D_{px}}{\text{pixels per mm}} \tag{10}$$

The conversion from pixels to millimeters was based on the reference marker in each image and set to 9 pixels/mm. The ABCD rule was then used to break the lesion diameter down into ordinal categories. Scores $D \in [1, 6]$ were given based on D_{mm}, with $D = 1$ for $D_{mm} \leq 1$ and $D = 6$ for $D_{mm} > 5$, and so on for the middle intervals [16].

2.5 Rule Scoring

The TDS was computed as a weighted sum of the determined ABCD scores:

$$TDS = 1.3 \cdot A + 0.1 \cdot B + 0.5 \cdot C + 0.5 \cdot D \tag{11}$$

These weights were adopted from the standard ABCD rule formulation in dermatology, and the final classification was defined as NV if $TDS < 4.75$, DN if $4.75 \leq TDS \leq 5.75$, and MM if $TDS > 5.75$ [9].

3 Results

To be able to localize the lesion and anonymize the data, a cropping step was performed using the YOLOv8n object detection model. YOLOv8n was trained for 20 epochs with an 80/20 split for training and validation steps. The performance was evaluated using mAP_{5095}, mAP_{50}, mAP_{75}, mean F1-score, and per-class average precision (AP). The obtained results are summarized in Table 2.

Table 2. Validation metrics of the YOLOv8n model

$\text{mAP}_5 0 - 95$	mAP_{50}	mAP_{75}	F1	AP
0.8220	0.9800	0.9063	0.9472	0.8220

The high mAP_{50} value of 0.98 reported in Table 2 indicates good localization capability, while the mAP_{5095} score of 0.8220 reflects robust performance across varying IoU thresholds. The F1-score of 0.9472 further confirms a favorable balance between precision and recall, demonstrating reliable detection of the reference marker.

The effect of the preprocessing pipeline was evaluated by analyzing the impact of illumination correction. Gray-world normalization was applied to reduce the overall color cast in the images. Subsequently, reference-based white balancing further improved color consistency by adjusting the RGB channels according to a detected white region.

Qualitative analysis shows that the suggested preprocessing pipeline (Fig. 4) makes the lesions look more consistent by eliminating color distortions that occur in different acquisition conditions. After correction, red and brown tones stand out more, and blue-gray areas are shown more clearly. This improvement makes it easier to classify colors accurately and then extract ABCD features.

3.1 Feature Analysis

The extracted ABCD features show consistent patterns across lesion categories. MM lesions exhibit the highest values across all components, while nevus NV lesions show the lowest, with DN falling in between.

Mean asymmetry increases from 0.388 for NV to 0.419 for DN and 0.567 for MM. Border irregularity follows a similar trend, with mean values of 1.167 (NV), 1.519 (DN), and 2.273 (MM). The number of detected colors also rises

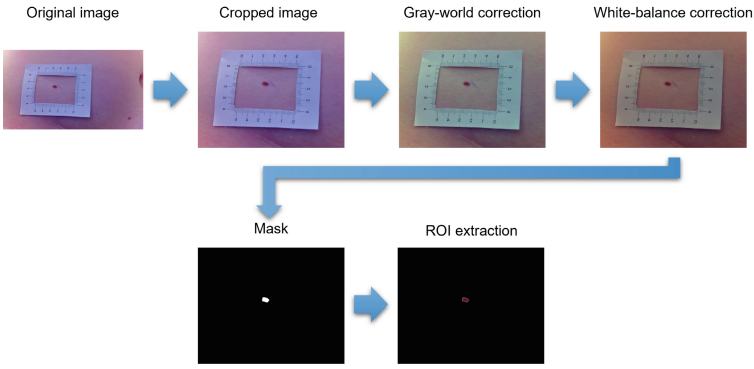


Fig. 4. Preprocessing pipeline.

progressively, from 2.778 (NV) to 3.538 (DN) and 4.045 (MM). Diameter shows the most pronounced separation, with melanoma reaching a mean of 7.995 mm, compared to 4.715 mm for DN and 2.723 mm for NV (Table 3).

Table 3. Mean values for the features extracted for each lesion class

Class	A	B	C	D
NV	0.388	1.167	2.778	2.723
DN	0.419	1.519	3.538	4.715
MM	0.567	2.273	4.045	7.995

To further validate the proposed framework, a qualitative analysis of the extracted ABCD features was performed. These examples highlight the ability of the method to capture clinically relevant characteristics across different lesion types.

Asymmetry. The asymmetry was analyzed by measuring how much the mirrored parts of the lesion overlapped along both the vertical and horizontal axes. NV lesions are usually symmetrical, with a lot of overlap between the two halves. On the other hand, MM lesions show clearer asymmetry, which is shown by less overlap and a more uneven distribution. DN possesses traits that are between benign and malignant patterns, as shown in Fig. 5.

Border Irregularity. Border irregularity was assessed by looking at both convexity deviation and radial distance variation. As demonstrated in Fig. 6, benign lesions usually have smooth, well-defined edges that are quite close to their convex hull. Melanoma lesions, on the other hand, have contours that are uneven, with big changes from convexity and more radial variance, which matches the real-life clinical scenarios.

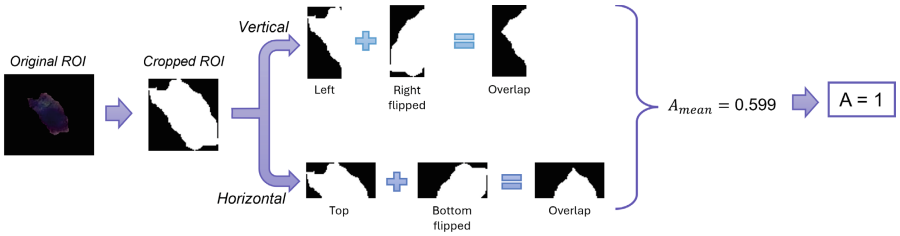


Fig. 5. Asymmetry score computed for a DN.

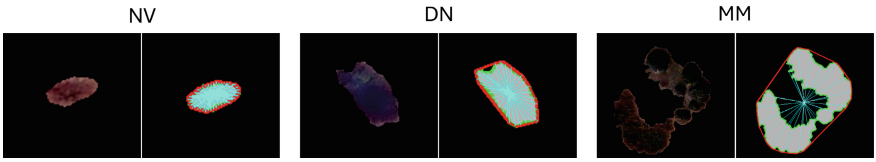


Fig. 6. Visualization of border irregularity for NV, DN, and MM.

Color Analysis. Color analysis confirms that melanoma lesions exhibit significantly higher chromatic heterogeneity compared to benign nevi, as seen in Table 4. The average number of detected colors increases from 2.778 (NV) to 4.045 (MM), reflecting the increased structural complexity of malignant lesions. Additionally, color uniformity decreases with lesion severity. NV exhibits higher uniformity, while MM shows the lowest values, indicating a more heterogeneous color distribution.

Table 4. Color statistics across lesion classes

Class	Mean Colors	Mean Uniformity
NV	2.778	0.681
DN	3.538	0.519
MM	4.045	0.364

Since the images were acquired under non-controlled conditions using a standard RGB camera and without polarized light, as in dermoscopic imaging, the reference RGB triplets and thresholds presented in Table 1 were not directly suitable for our dataset. Variations in illumination, reflections, and sensor response significantly affect color representation.

To address this limitation, a data-driven refinement of the color model was performed. Color quantization revealed that standard dermoscopic RGB triplets do not accurately represent the color distribution of non-dermoscopic images. In particular, blue-gray regions were often underrepresented, while red tones were frequently confused with brown shades (Fig. 7) due to illumination variability.

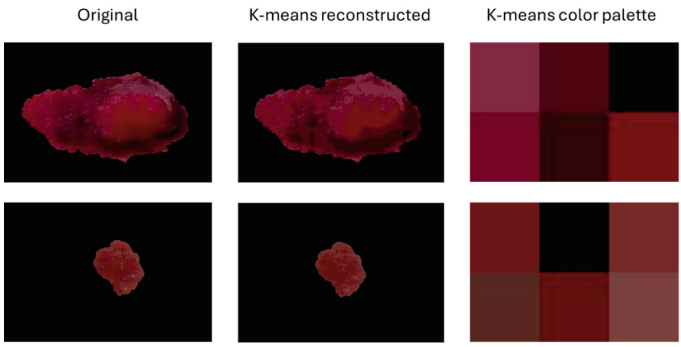


Fig. 7. Red color palette obtained through color quantization applied on the dataset. (Color figure online)

By applying K-means clustering to lesion pixels, dataset-specific color centroids were obtained, allowing refinement of reference triplets and thresholds. This step improved the consistency of color classification and enhanced the detection of clinically relevant features such as blue-gray structures. In Fig. 8, we show the 6 predominant colors in the entire dataset, lesions containing blue-gray or red components.

Fig. 8. Extracted color palette from the dataset, including blue-gray and red subsets. (Color figure online)

After determining the improved RGB triplets, we changed the reference and the threshold values for the following colors: red, light brown, and blue-gray. The improved RGB triplets can be seen in Fig. 9.

The hybrid color classification approach improved robustness by combining rule-based and distance-based methods, where the rule-based stage captured well-defined colors such as red and brown, and the distance-based stage ensured complete pixel assignment. An example is shown in Fig. 10, where blue-gray

Fig. 9. Comparison between dermoscopic reference color triplets and dataset-adapted RGB triplets.

regions were not detected using the original RGB triplets but were successfully identified after applying the proposed hybrid approach.

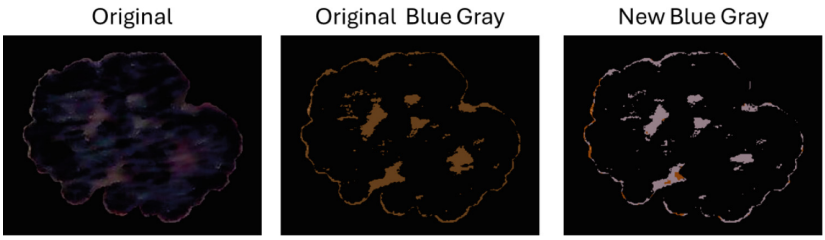


Fig. 10. Example of improved blue-gray detection after color quantization. (Color figure online)

The introduction of a relative noise threshold improved color stability by removing insignificant color components. Lower thresholds suppressed clinically relevant minor colors, whereas refined thresholds preserved diagnostically meaningful variations. As shown in Table 5, a threshold of 0.5 yielded the highest classification accuracy and was therefore adopted in the final hybrid color classification.

Table 5. Effect of noise threshold on classification performance

Threshold (%)	0	0.3	0.5	0.8
Accuracy (%)	68.49	69.18	70.55	68.49

3.2 Diameter

Diameter was estimated using the reference marker to convert pixel measurements into physical units. As illustrated in Fig. 11, melanoma lesions tend to

have larger diameters compared to NV, consistent with clinical guidelines. The use of a physical calibration marker ensures accurate and consistent measurement across all images.

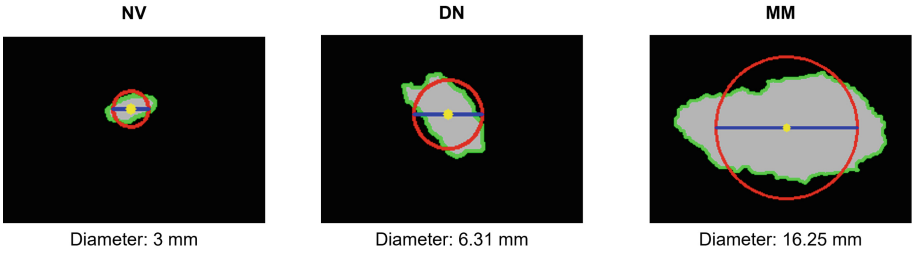


Fig. 11. Example of diameter computation.

3.3 Overall ABCD-Based Classification Performance

The proposed framework was evaluated on a dataset of 146 lesions, including 72 NV, 47 DN, and 27 MM cases. Using the computed ABCD features and the TDS, the method achieved an overall classification accuracy of 70.55%, with 103 correctly classified cases. A class-wise analysis shows that NV were classified with the highest accuracy, achieving 90.2%. MM cases reached a recall of 59.26%, while DN remained the most challenging class, with a recall of 46.81%. This behavior is consistent with clinical expectations, as DN represent an intermediate category with overlapping features between benign and malignant lesions, making them inherently more difficult to classify using rule-based approaches (Table 6).

Table 6. Classification performance per lesion type

	MM	DN	NV
Correct	16	22	65
Incorrect	11	25	7

To further evaluate the discriminative ability of the proposed framework, the continuous TDS values were analyzed using a ROC curve in a one-vs-rest setting, where melanoma cases were treated as the positive class and all remaining lesions as the negative class, as seen in Fig. 12. The ROC analysis showed an area under the curve (AUC) of 0.8358, which means that the TDS worked well at distinguishing the difference between images that were not dermoscopic. Youden's index showed that the best decision threshold was 5.7, which is similar to the threshold for MM detection that is accepted clinically. The method achieved 62.96% sensitivity and 94.12% specificity, revealing that the proposed method

can correctly determine non-melanoma lesions. These results show that the proposed ABCD-based framework performs well to keep the TDS's diagnostic value, even when used with standard RGB images.

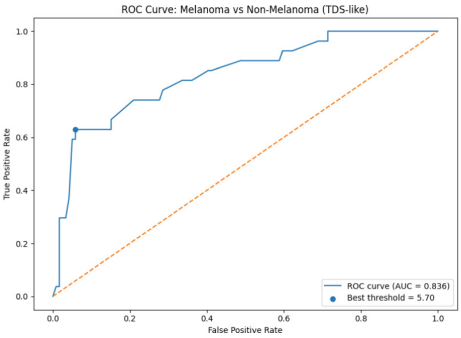


Fig. 12. ROC curve for MM versus non-MM classification using TDS.

Figure 13 shows how TDS values are dispersed across different types of lesions. The classes can be clearly separated because the mean TDS values are 3.42 for NV, 4.56 for DN, and 5.57 for MM. This gradual rise is expected from the ABCD rule in a clinical setting, where malignant lesions show more asymmetry, border irregularity, color variation, and diameter. The fact that DN overlaps with the other classes helps explain why this category had a lower classification performance.

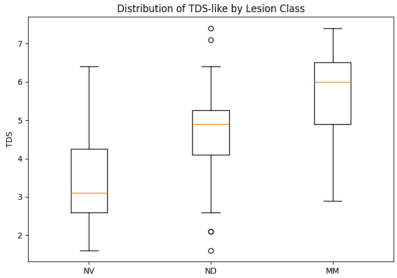


Fig. 13. TDS values distribution across lesion classes.

The suggested score does not seek to duplicate dermoscopic performance but to maintain clinically significant trends. In our dataset, all ABCD-derived components and the final score increase from NV to DN and MM. This shows that the adapted framework preserves the diagnostic structure in non-dermoscopic RGB images. The primary ambiguity occurs within the dysplastic category, aligning with its intermediate clinical characteristics.

These results suggest that ABCD-like features can provide useful support for dermatologists even in the absence of dermoscopic detail. However, the resulting

score is not directly equivalent to the dermoscopic TDS, as it is derived under different acquisition conditions and relies on more limited visible structures. Future work should focus on larger datasets, evaluating robustness under varying lighting conditions, and comparing the approach with both dermoscopic and learning-based methods.

4 Conclusions

This work presented a clinically-inspired approximation of ABCD feature extraction for non-dermoscopic RGB images acquired in realistic clinical conditions. Rather than directly reproducing dermoscopic TDS, the proposed framework investigates whether ABCD-derived descriptors can preserve diagnostically relevant trends when applied to standard visible-spectrum imaging. The results show that the adapted score captures meaningful differences between benign and malignant lesions, while also highlighting the difficulty of the dysplastic category. These findings support the potential of interpretable, low-cost RGB-based analysis as a complementary direction for melanoma screening in settings where dermoscopic imaging is not consistently available.

Acknowledgements. This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCDI-UEFISCDI, project number PN-IV-P7-7.1-PED-2024-1873, within PNCDI IV. The authors sincerely thank the hospitals and dermatologists for granting them access to acquire the necessary images for this study. They also extend their gratitude to all the volunteers who participated in this research and contributed to their work.

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

References

1. Cavalcanti, P.G., Scharcanski, J., Baranoski, G.V.: A two-stage approach for discriminating melanocytic skin lesions using standard cameras. Expert Syst. Appl. **40**(10), 4054–4064 (2013)
2. Chatterjee, S., Dey, D., Munshi, S., Gorai, S.: Dermatological expert system implementing the ABCD rule of dermoscopy for skin disease identification. Expert Syst. Appl. **167**, 114204 (2021)
3. Hofman, M.S., Hicks, R.J.: Role of pet/ct in melanoma. In: PET/CT in melanoma, pp. 37–51. Springer (2017)
4. International Agency for Research on Cancer, World Health Organization: Global burden of cutaneous melanoma in 2020 and projections to 2040 (May 2022). https://www.iarc.who.int/infographics/global-burden-of-cutaneous-melanoma-in-2020-and-projections-to-2040/, Accessed 30 March 2026
5. Jartarkar, S.R., et al.: New diagnostic and imaging technologies in dermatology. J. Cosmet. Dermatol. **20**(12), 3782–3787 (2021)

6. Kotla, H., Rajasekaran, A.K., Rana, H.: A multi-task deep learning framework for skin lesion classification, ABCDE feature quantification, and evolution simulation, arXiv preprint arXiv:2510.14855 (2025)
7. Mabrouk, M.S., Sayed, A.Y., Afifi, H.M., Sheha, M.A., Sharwy, A.: Fully automated approach for early detection of pigmented skin lesion diagnosis using ABCD. J. Healthcare Inform. Res. **4**(2), 151–173 (2020)
8. Messadi, M., Cherifi, H., Bessaid, A.: Segmentation and ABCD rule extraction for skin tumors classification, arXiv preprint arXiv:2106.04372 (2021)
9. Pathan, S., Prabhu, K.G., Siddalingaswamy, P.: Techniques and algorithms for computer aided diagnosis of pigmented skin lesions–a review. Biomed. Signal Process. Control **39**, 237–262 (2018)
10. Pelagotti, A., Ferrara, P., Pescitelli, L., Gerlini, G., Piva, A., Borgognoni, L.: Non-invasive inspection of skin lesions via multispectral imaging. In: Optical Methods for Inspection, Characterization, and Imaging of Biomaterials, vol. 8792, pp. 55–67. SPIE (2013)
11. Roky, A.H., et al.: Overview of skin cancer types and prevalence rates across continents. Cancer Pathogenesis Therapy **3**(02), 89–100 (2025)
12. Senan, E.M., Jadhav, M.E.: Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer. Global Transitions Proc. **2**(1), 1–7 (2021)
13. Sultana, A.E., Oniga, M., Rus, P.F., Dobre, A.L., Orzan, O.A.: Melanodet: multimodal imaging system for screening and analysis of cutaneous melanoma. IEEE Access (2025)
14. Thompson, J.F., Scolyer, R.A., Kefford, R.F.: Cutaneous melanoma. The Lancet **365**(9460), 687–701 (2005)
15. Yu, Z., et al.: Early melanoma diagnosis with sequential dermoscopic images. IEEE Trans. Med. Imaging **41**(3), 633–646 (2021)
16. Zaqout, I.: Diagnosis of skin lesions based on dermoscopic images using image processing techniques. Pattern Recognition-Selected Methods and Applications. IntechOpen (2019)

LEAP: A Leakage-Free Evolutionary Alpha Pipeline for NLP-Driven DJIA Prediction

Miguel P. Cunha[1(✉)], Bruno L. Dalmazo[1], Viviane Leite Dias de Mattos[1], Richard F. Pinto[1], Diego Renan Bruno[2(✉)], Eduardo N. Borges[1], Giancarlo Lucca[1], Fabian C. Cardoso[3], and Rafael A. Berri[1]

[1] Federal University of Rio Grande (FURG), Rio Grande, Brazil
{miguel.pinho,dalmazo,vivianemattos,richard_pinto,eduardoborges,giancarlo.lucca,rafaelberri}@furg.br
[2] São Paulo State University (UNESP), São José do Rio Preto, Brazil
diego.bruno@unesp.br
[3] University of Rio Verde (UniRV), Rio Verde, Brazil
fabian@unirv.edu.br

Abstract. Financial market prediction utilizing Natural Language Processing (NLP) is notoriously challenging due to low signal-to-noise ratios, high dimensionality, and the persistent flaw of evaluating trading models using symmetric statistical metrics like Accuracy or Area Under Receiver Operating Characteristic Curve (AUC-ROC). To deal with these questions, we propose in this paper a rigorous data-leakage-free machine learning pipeline to forecast the daily directional movement of the Dow Jones Industrial Average (DJIA). Our architecture integrates FinBERT-derived sentiment embeddings with technical indicators and temporal lags. To combat overfitting, we employ Recursive Feature Elimination with Cross-Validation (RFECV) for dimensionality reduction and a Genetic Algorithm (GA) to optimize hyperparameters by explicitly penalizing the generalization gap. By applying Cost-Sensitive Learning to prioritize high-volatility "fat-tail" events, we contrast a Gradient Boosting ensemble with a Multi-Layer Perceptron (MLP) via an out-of-sample financial backtest. These findings point to a statistical-financial paradox: despite the Gradient Boosting's better discrimination (Test AUC 0.5041, Alpha +1.69 p.p.), the MLP statistically "fails" (Test AUC 0.4812) yet takes advantage of market asymmetries to generate a +13.27 p.p. Alpha. These findings indicate that symmetric evaluation metrics are not enough for assessing trading viability, pointing to the need for magnitude-aware training.

Keywords: Algorithmic Trading · Natural Language Processing · Genetic Algorithms · Feature Selection · Financial Time Series

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 71–87, 2027.
https://doi.org/10.1007/978-3-032-30494-0_5

1 Introduction

The prediction of financial market movements is a known difficult problem in computational finance [14]. Financial time series are inherently noisy, non-stationary, and heavily influenced by stochastic external events and human behavioral biases. In recent years, the integration of Natural Language Processing (NLP) into algorithmic trading has brought relevant gains to the field. By extracting sentiment signals from financial news, models can quantify market psychology before it is fully priced into historical data [16]. However, the combination of textual embeddings with traditional technical indicators raises the dimensionality of the feature space, increasing the risk of model overfitting, a known limitation in predictive finance [2].

A recurring issue in the applied Machine Learning literature for financial forecasting is the reliance on symmetric statistical metrics, such as Accuracy or AUC-ROC, as the main criteria for model evaluation [10]. These metrics tend to assume that all misclassifications carry the same cost. In financial markets, this assumption is not valid due to the presence of "fat tails" (high-magnitude, low-frequency price movements). A model can fail to predict the market direction on several low-volatility days, degrading its AUC, but still correctly anticipate a market crash or a strong rally, while producing relevant trading gains [13].

To address these challenges, this paper proposes LEAP (Leakage-Free Evolutionary Alpha Pipeline), a data-leakage-free optimization pipeline for predicting the daily directional movement of the Dow Jones Industrial Average (DJIA). Our approach combines domain-specific NLP sentiment (extracted via FinBERT) with classical technical indicators [4,15] and multi-scale temporal lags. To deal with high dimensionality and avoid noise memorization, we employ Recursive Feature Elimination with Cross-Validation (RFECV) combined with an Evolutionary Optimization framework (Genetic Algorithms) [1]. The genetic algorithm's fitness function was built to penalize the generalization gap between training and validation sets, pushing the search toward more generalizable models.

We also apply a Cost-Sensitive Learning approach that uses sample weights based on return magnitude, guiding the models to focus on high-volatility days. Two learning architectures are compared: a Gradient Boosting ensemble and a Multi-Layer Perceptron (MLP) neural network. The evaluation goes beyond standard statistical metrics by using an out-of-sample financial backtest with a dynamic decision threshold.

The primary contributions of this work are three-fold:

1. **A Leakage-Free Predictive Pipeline:** We introduce a highly disciplined chronological framework that integrates NLP sentiment and temporal data while strictly preventing forward-looking bias during feature selection and normalization.
2. **Evolutionary Overfitting Control:** Evolutionary Overfitting Control is applied to the Genetic Algorithm for optimizing the objective function and successfully selecting hyperparameters in highly stochastic environments.

3. **Empirical Demonstration of the Statistical-Financial Paradox:** We reveal that symmetric metrics are insufficient for trading viability. Our results show that while the Gradient Boosting model maintains statistical stability (AUC >0.50) yielding a $+$**1.69 p.p. Alpha**, the MLP architecture, despite a seemingly failing statistical metric (AUC <0.50), effectively captures market asymmetries, delivering a massive $+$**13.27 p.p. Alpha**.

The remainder of this paper is organized as follows: Sect. 2 outlines the theoretical foundations, followed by a review of related literature in Sect. 3. Section 4 details the proposed LEAP methodology. Section 5 discusses the empirical results and the financial backtesting paradox. Finally, Sect. 6 concludes the study.

2 Theoretical Background

This study is built around two learning architectures: Gradient Boosting (GB) and the Multi-Layer Perceptron (MLP). GB is a sequential tree-based ensemble that minimizes differentiable loss, known for its good results on noisy tabular data [7]. The MLP is a feed-forward artificial neural network able to handle non-linear interactions without relying on rigid decision boundaries. To optimize these architectures, we employ Genetic Algorithms (GA) [1], an evolutionary search heuristic that searches the hyperparameter space more effectively than grid-based methods by simulating natural selection. In quantitative finance, models are typically assessed using symmetric statistical metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). However, financial returns exhibit leptokurtic distributions characterized by "fat tails" [8]. Consequently, symmetric metrics treat all prediction errors the same way, leaving out the economic impact of asymmetric returns. To address this, Cost-Sensitive Learning introduces magnitude-based sample weights into the training phase, making the loss function give more weight to errors on high-volatility days than on low-magnitude ones.

3 Related Work

Financial forecasting puts pressure on the Efficient Market Hypothesis (EMH) [9] by taking advantage of cognitive biases and momentum effects [12]. In recent years, Natural Language Processing (NLP) has become a useful resource for capturing these behavioral shifts. Transformer-based architectures, specifically FinBERT [16], have shown better results than generic models when handling domain-specific financial terms. However, integrating high-frequency news sentiment with technical indicators brings collinearity problems and higher dimensionality to the feature space. To deal with this, previous works point to feature selection techniques such as Recursive Feature Elimination with Cross-Validation (RFECV) [11].

A common problem in applied financial machine learning is the heavy dependence on symmetric metrics as the main measure of model quality [10]. Models

tend to focus too much on frequent, low-magnitude days while overlooking less frequent directional changes that produce real trading gains [13]. While existing literature explores hyperparameter tuning to improve accuracy, traditional methods like Grid Search are computationally costly and likely to overfit in non-stationary settings [3]. This work focuses on filling this gap by introducing a leakage-free pipeline that combines NLP embeddings, Cost-Sensitive Learning, and evolutionary regularization to directly account for the generalization gap and focus on out-of-sample financial performance rather than in-sample symmetric metrics.

4 Methodology

Our methodology follows a sequential, leakage-free pipeline to forecast the DJIA's daily direction. To build this framework, we merge multi-scale temporal lags, classical technical indicators, and NLP-based sentiment signals. Figure 1 details this chronological setup. We execute all feature construction, normalization, and selection strictly within the in-sample data, ensuring no future information leaks prior to the out-of-sample evaluation. Finally, a financial backtest assesses the model's cumulative returns on the held-out test set. This step verifies the system's practical trading potential, giving a more grounded view than standard classification metrics alone.

4.1 Data Inputs

Our study spans August 2008 to July 2016, pairing two public datasets. To capture macro-level psychology rather than direct financial news, we extracted daily geopolitical headlines from the *r/worldnews* board using the *Combined News DJIA* collection. These sentiment signals were linked to standard OHLCV metrics from the *DJIA Historical Prices* dataset. Aligning both sources by their `Date` keys and filtering out incomplete records resulted in a continuous series of 1,989 trading days.

The prediction target is the binary *directional movement* of the DJIA on the following trading day:

$$\text{Label}_t = \begin{cases} 1 & \text{if } R_{t+1} > 0 \quad \text{(market rises)}, \\ 0 & \text{otherwise}, \end{cases} \tag{1}$$

where $R_{t+1} = (C_{t+1} - C_t)/C_t$ is the simple next-day return and C_t denotes the closing price at day t.

To reflect the economic significance of each trading day, a *sample-weighting* scheme is applied. Days whose absolute return exceeds the 75^{th} percentile of the return-magnitude distribution receive a weight of 2.0; all remaining days receive a weight of 1.0. This Pareto-inspired threshold captures the top 25% of moves most likely to exceed transaction costs and generate exploitable alpha [13], while keeping the class distribution well-balanced (≈52% up/48% down), precluding the need for resampling.

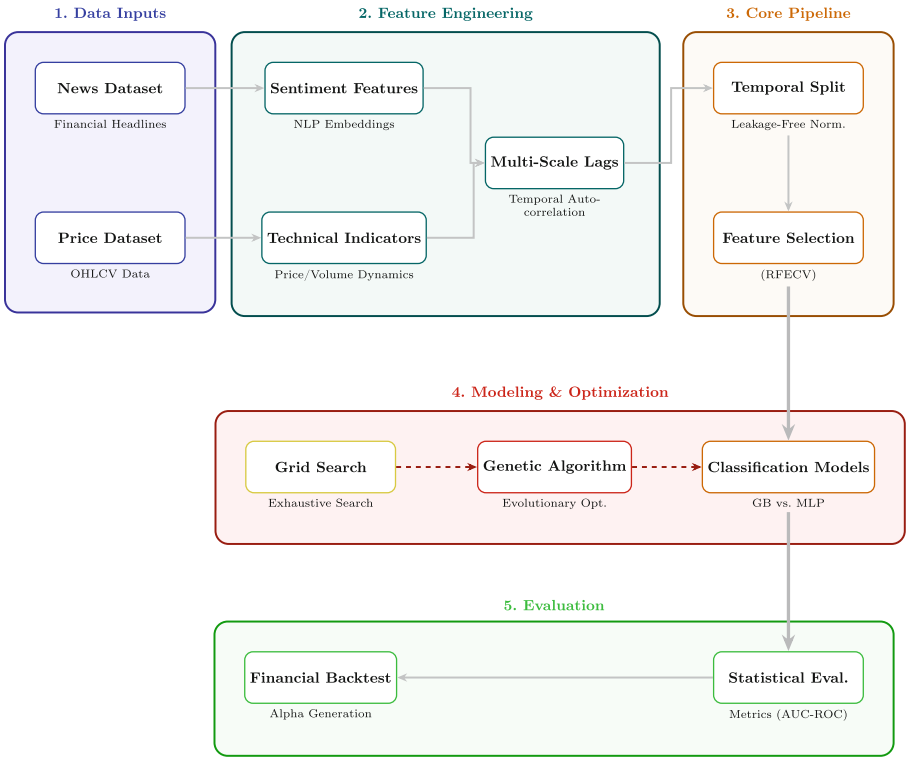


Fig. 1. Overview of our Leakage-Free Evolutionary Alpha pipeline with a two-layered module architecture. The upper layer flows horizontally from raw datasets through feature engineering to the core processing pipeline. A vertical main arrow connects to the lower layer, which integrates model training, comparative hyperparameter optimization, and final statistical and financial performance evaluation.

4.2 Feature Engineering

Raw inputs are transformed into a rich, multi-modal feature space spanning three complementary information sources: NLP-derived sentiment from news headlines, price- and volume-based technical indicators, and multi-scale temporal lags that encode autocorrelation structure. Together they produce **109 candidate features**.

Sentiment Features via FinBERT. Financial news headlines carry a substantial predictive signal but require a domain-adapted language model to properly decode financial vocabulary and tone. The extraction of sentiment signals was performed using the *FinBERT-Tone* model[1], a BERT-based architecture. While more recent large language models (LLMs) and generative architectures

[1] Available at: https://huggingface.co/yiyanghkust/finbert-tone.

have emerged since 2020, we opted for *FinBERT* due to its specialized fine-tuning on financial corpora and its superior computational efficiency. This balance is critical for an evolutionary pipeline, where sentiment extraction must be integrated into an optimization loop without prohibitive latency. Furthermore, as our dataset spans 2008 to 2016, *FinBERT* effectively captures the language patterns of this period without requiring the excessive world knowledge of larger generative models. This resolves the ambiguity of terms such as "liability" or "share", which have neutral connotations in common English but carry strong semantic weight in a market context.

Given a headline H, the model processes the token sequence and outputs logits for three classes: Positive (z_{pos}), Negative (z_{neg}), and Neutral (z_{neu}). We compute the likelihood of each class by passing the model's output through a Softmax layer. This step gives us the following probabilities:

$$P(c) = \frac{e^{z_c}}{\sum_{k \in \{pos,neg,neu\}} e^{z_k}} \tag{2}$$

Working with three separate dimensions per news item complicates the feature space. Therefore, we compress these outputs into a *Compound Score* (S). This metric maps the headline's mood onto a $[-1, 1]$ scale by taking the difference between the positive and negative likelihoods:

$$S = P(pos) - P(neg) \tag{3}$$

The original dataset provides 25 headlines per day ($H_1, \ldots, H_{25}$). To convert this text corpus into a structured format suitable for supervised classification models, we ran 26 statistical transformations grouped into four categories: Central Tendency, Distribution, Raw Probabilities, and Temporal Dynamics. One contribution of this work is the *Weighted Sentiment*, which gives lower weights to headlines based on their position in the daily ranking, reflecting the idea that investors tend to pay more attention to top-ranked news. The weight w_i of headline i is given by:

$$w_i = e^{-0.3 \cdot (i-1)} \tag{4}$$

Table 1 presents a comprehensive description of all extracted features. All temporal features are computed strictly in chronological order, ensuring no future information contaminates the feature matrix.

Technical Indicators. To complement the textual signal with price-based market dynamics, we compute **39 technical indicators** across five categories, with all price-derived quantities expressed as normalised ratios to ensure stationarity and cross-period comparability:

– **Trend** (13): Close-to-SMA ratios for windows $\{5, 20, 50, 100\}$; Close-to-EMA ratios for windows $\{12, 26\}$; SMA crossover ratios (5/20, 20/50, 50/100); binary signals for price above SMA_{20}, SMA_{50}, SMA_{100}, and a fast golden cross ($\text{SMA}_{20} > \text{SMA}_{50}$).

Table 1. Comprehensive Description of NLP Sentiment Features

Feature (Code)	Calculation Description	Academic Justification
1. Central Tendency and Weighting		
sent_mean_compound	Arithmetic mean of S scores.	Daily sentiment baseline (equal weighting).
sent_median_compound	Median of S scores.	Robustness against outliers and isolated extreme headlines.
sent_weighted_mean	$\frac{\sum(S_i \cdot w_i)}{\sum w_i}$	Simulates attention bias (top headlines carry more weight).
sent_weighted_std	Standard deviation weighted by w_i.	Measures opinion divergence in major news.
sent_std_compound	Simple standard deviation.	Indicator of ambiguity and noise in the news flow.
sent_min_compound	Lowest S value of the day.	Tail risk detector (Panic/FUD).
sent_max_compound	Highest S value of the day.	Euphoria detector (FOMO).
2. Raw Probabilities		
sent_mean_pos	Mean of $P(pos)$ probabilities.	Isolates the pure optimism signal.
sent_mean_neg	Mean of $P(neg)$ probabilities.	Isolates the fear signal (Loss Aversion).
sent_mean_neu	Mean of $P(neu)$ probabilities.	Indicates the absence of fundamental drivers.
3. Distribution and Shape		
sent_pct_positive	% of news with $S > 0.1$.	Market breadth (Generalized optimism).
sent_pct_negative	% of news with $S < -0.1$.	Breadth of pessimism.
sent_pct_neutral	% of news between -0.1 and 0.1.	Market inertia and indecision.
sent_dominance	S value with the highest absolute $\|S\|$.	Identifies if the day is driven by a strong single event.
sent_consistency	$1/(1+\sigma_s)$.	Signal confidence (Consensus among sources).
sent_skewness	Skewness of the S distribution.	Detects hidden risks on apparently neutral days.
sent_range_norm	$S_{max} - S_{min}$.	Polarization and narrative conflict.
4. Temporal Dynamics (Time Series)		
sent_mean_change	$S_t - S_{t-1}$ (1st Derivative).	Captures the change in expectations (Delta).
sent_mean_change_pct	$(S_t - S_{t-1})/\|S_{t-1}\|$	Percentage change in mood (Relative sensitivity).
sent_acceleration	$\Delta S_t - \Delta S_{t-1}$ (2nd Derivative).	Strength of trend reversal.
sent_momentum	Rolling sum of S (3 d).	Recent cumulative psychological effect.
sent_volatility_5d	Rolling standard deviation (5 d).	Market uncertainty (Weekly).
sent_volatility_10d	Rolling standard deviation (10 d).	Market uncertainty (Bi-weekly).
sent_vs_sma5	$S_t - \text{SMA}_5(S)$.	Mean Reversion indicator (Short-term).
sent_vs_sma10	$S_t - \text{SMA}_{10}(S)$.	Mean Reversion indicator (Medium-term).

- **Momentum** (11): RSI_{14} with overbought/oversold flags; MACD, signal line, and histogram normalised by closing price with a bullish crossover flag; ROC_{12}; Stochastic Oscillator (K, D, bullish flag).
- **Volatility** (7): Bollinger Band upper and lower boundaries normalised by closing price; band width; relative position within the band $\in [0, 1]$; above-upper and below-lower breakout flags; ATR_{14} normalised by closing price.
- **Volume** (4): OBV percentage change; 20-day volume ratio; high-volume binary flag (ratio > 1.5); MFI_{14}.
- **Trend strength** (4): ADX_{14}, positive DI, negative DI, and a strong-trend flag (ADX > 25).

Multi-scale Lag Features. To explicitly encode short- and medium-term autocorrelation structure [12], we augment the feature matrix with lagged OHLCV values and lagged daily returns across three temporal scales:

- **Ultra-short term** ($\ell \in \{1, 2\}$ days): captures mean-reversion signals immediately following large price moves.

- **Short term** ($\ell \in \{3, 4, 5\}$ days): encodes weekly momentum consistent with documented serial autocorrelation in equity returns.
- **Medium term** ($\ell \in \{10, 20\}$ days): reflects bi-weekly and monthly market cycles.

Applying the seven lag values to the five OHLCV series and to the daily return series yields **42 lag features**. Rows containing NaN values introduced by lagging are removed before any model fitting, preserving strict chronological integrity.

4.3 Core Pipeline

Temporal Split and Leakage-Free Normalization. Financial time series exhibit strong serial dependence, making standard random K-Fold cross-validation inappropriate [6]. We adopt a strict chronological partitioning of the 1,989 aligned trading days:

- **Training set (70%)**: used exclusively for model fitting, scaler calibration, and RFECV feature selection.
- **Validation set (15%)**: used for model selection and hyperparameter tuning.
- **Test set (15%)**: held out until final evaluation; never exposed to any fitting procedure, corresponding to the period from May 2015 to June 2016.

Normalization is performed using a `StandardScaler` fitted *exclusively* on the training partition. The learned statistics (μ_{train}, σ_{train}) are then applied to transform the validation and test sets:

$$x' = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}. \tag{5}$$

This design simulates a realistic deployment scenario in which future price statistics are unknown at prediction time, thereby eliminating a common but subtle source of data leakage [13].

Feature Selection via RFECV. With 109 candidate features, irrelevant and redundant inputs risk degrading generalisation through the curse of dimensionality. We apply Recursive Feature Elimination with Cross-Validation (RFECV) [11] exclusively within the training set, using a 5-fold time-series-aware scheme that preserves chronological order across folds.

Three estimators serve as wrapper models: Logistic Regression, Random Forest, and Gradient Boosting. Each wrapper eliminates features one at a time (step = 1) and selects the minimal subset that maximises 5-fold cross-validated accuracy on the training partition. The wrapper with the best validation-set accuracy sets the final feature subset shared across all classifiers, so that performance differences come from the learners themselves and not from differences in the input data. To run a fair comparison against the tree-based ensemble, the architecture goes through hyperparameter tuning via Genetic Algorithms

(GA). Unlike Grid Search, which is computationally costly and tends to overfit, the evolutionary approach uses natural selection to explore the hyperparameter search space. To reduce the risk of overfitting in financial time series, the GA relies on a penalized fitness function. Instead of only maximizing the cross-validated validation metric, the objective function penalizes the Generalization Gap, which is the difference between the training performance and the validation performance. Here, $\boldsymbol{\theta}$ stands for the hyperparameter vector of a given individual in the population. The fitness function $F(\boldsymbol{\theta})$ is defined as:

$$F(\boldsymbol{\theta}) = \mathrm{AUC}_{\mathrm{val}}(\boldsymbol{\theta}) - \lambda \cdot \max\left(0, \mathrm{AUC}_{\mathrm{train}}(\boldsymbol{\theta}) - \mathrm{AUC}_{\mathrm{val}}(\boldsymbol{\theta})\right). \tag{6}$$

In this formula, $\mathrm{AUC}_{\mathrm{train}}(\boldsymbol{\theta})$ and $\mathrm{AUC}_{\mathrm{val}}(\boldsymbol{\theta})$ represent the average ROC-AUC scores computed during the training and validation stages of the TimeSeriesSplit. The λ parameter controls the strength of this regularization. We used a value of 0.5 for λ throughout our tests. This choice actively filters out candidate models that overfit the historical data while failing to generalize to new data.

4.4 Modeling and Optimization

We evaluate a diverse set of classifiers representing distinct families of inductive bias. This comparative approach ensures that the performance gains derived from the feature engineering pipeline are rigorously validated across fundamentally different learning paradigms:

- **Logistic Regression (LR)**: A linear baseline providing interpretable coefficients and a lower-bound performance reference.
- **Random Forest (RF)**: A bagging ensemble providing variance reduction through feature randomization [5], offering inherent robustness to noise.
- **Gradient Boosting (GB)**: A sequential additive ensemble that minimizes a differentiable loss through boosting. Due to its recognized superiority in modeling noisy tabular data [7], GB is treated as our primary tree-based architecture.
- **Multi-Layer Perceptron (MLP)**: A feed-forward neural network utilized not merely as a comparative baseline, but to explicitly capture highly non-linear, asymmetric interactions between NLP sentiment shocks and technical dynamics that rigid tree-based splits might fail to isolate.

All models are configured with `class_weight='balanced'` to compensate for any residual class imbalance in the training folds. Crucially, during the training phase, the magnitude-based sample-weight vector defined in Eq. 1 is passed to the objective function. This Cost-Sensitive Learning approach directs the gradient descent to penalize misclassifications on economically significant days more heavily.

Evolutionary Hyperparameter Optimization. Financial time series tend to overfit easily, and using default hyperparameters can create a large Generalization Gap between in-sample memorization and out-of-sample performance. To handle this, both core architectures (Gradient Boosting and MLP) go through a careful hyperparameter tuning process. Traditional methods, such as Grid Search, tend to reach overfitted solutions by simply maximizing validation metrics without considering the noise present in financial data. To move past this limitation, we use an Evolutionary Optimization approach via Genetic Algorithms (GA) combined with a chronological *TimeSeriesSplit* cross-validation scheme. The evolutionary process relies on a penalized fitness function that accounts for the generalization gap, pushing the algorithm to drop configurations that fit the training data well but struggle on validation. For the Gradient Boosting model, the GA searches through the regularization space, tuning structural parameters such as maximum tree depth, learning rate, and subsample ratios. For the MLP architecture, the GA looks into the network topology space, tuning hidden layer dimensions, L_2 regularization limits, learning rates, and solver dynamics. This approach helps the final models find a balance between predictive power and generalization needed to produce steady out-of-sample financial gains.

4.5 Evaluation

Accuracy and AUC-ROC measure statistical discrimination but do not directly quantify whether a model generates exploitable profit in a realistic trading setting. We therefore complement all classification metrics with a **Financial Backtest** that translates predictions into cumulative returns, which constitutes the primary evaluation criterion of this work.

The strategy adopts a *Trend Following* logic applied to the optimized models on the held-out test set. At each trading day t, the model outputs a probability $\hat{p}_t = P(\text{Label} = 1 \mid \mathbf{x}_t)$. The resulting market position is determined by a dynamic decision threshold τ:

$$\text{pos}_t = \begin{cases} +1 \text{ (long)} & \text{if } \hat{p}_t > \tau, \\ -1 \text{ (short)} & \text{otherwise.} \end{cases} \tag{7}$$

The daily strategy return is $r_t^s = \text{pos}_t \times R_{t+1}$, and the cumulative capital curve is:

$$W_T = \prod_{t=1}^{T} (1 + r_t^s), \quad W_0 = 1. \tag{8}$$

The benchmark is a passive Buy-and-Hold strategy, $W_T^{\text{BH}} = \prod_{t=1}^{T}(1+R_{t+1})$, and the excess return (alpha) over the test period is reported as $\alpha = W_T - W_T^{\text{BH}}$.

To prevent the strategy from being systematically biased toward long or short positions due to the inherent class probability shift of the evolved models, we do not rely on a fixed 0.50 decision boundary. Instead, we implement a *Dynamic Threshold* mechanism. The decision threshold τ is set to the median prediction

probability of the validation set, keeping market exposure balanced while avoiding look-ahead bias. The strategy works with unit position sizing and leaves transaction costs out, choices that fit the scope of this work, which focuses on measuring the raw predictive alpha of the pipeline.

5 Results and Discussion

In this section, we present the empirical results of the proposed methodology, divided into three fundamental stages: dimensionality reduction via RFECV, hyperparameter optimization guided by a Genetic Algorithm (GA), and finally, the out-of-sample performance evaluation contrasting traditional statistical metrics with actual financial returns (Backtest).

5.1 Feature Selection and Dimensionality Reduction

The integration of natural language processing embeddings (FinBERT) with traditional tabular variables generated an initial vector space of 109 features. Since tree-based models are susceptible to overfitting in high-dimensional spaces with inherent noise, the Recursive Feature Elimination with Cross-Validation (RFECV) method was applied.

Figure 2 illustrates the selection process for the Gradient Boosting model. The algorithm identified an optimal subset of only 21 features, achieving a Mean CV Accuracy of 0.5180. The green shaded area represents the standard deviation across the cross-validation folds, highlighting that the removal of 88 redundant variables not only reduced computational complexity but also stabilized the model's variance regarding unseen data.

5.2 Evolutionary Optimization of Hyperparameters

With the reduced feature space, the Gradient Boosting model was subjected to optimization via a Genetic Algorithm. To mitigate the memorization of training data (overfitting), the fitness function was modeled as a penalized metric, subtracting the Gap between the Train AUC and Validation AUC from the main metric, as detailed in Sect. 4.

Figure 3 presents the optimization profile. The Early Stopping mechanism was triggered at the 32nd generation, after 25 generations without significant improvements, proving the algorithm's mathematical convergence. The optimal individual achieved a converged fitness of 0.3809, balancing predictive capacity with strong structural regularization (n_estimators=100, max_depth=2, subsample=0.7).

To provide full transparency regarding the evolutionary process, Table 2 details the search space constraints and the configuration of the Genetic Algorithm, alongside the final optimized hyperparameters obtained for the winning Gradient Boosting model. The algorithm successfully identified an optimal configuration characterized by strong structural regularization (e.g., shallow trees with `max_depth`=2 and a conservative `learning_rate`=0.01).

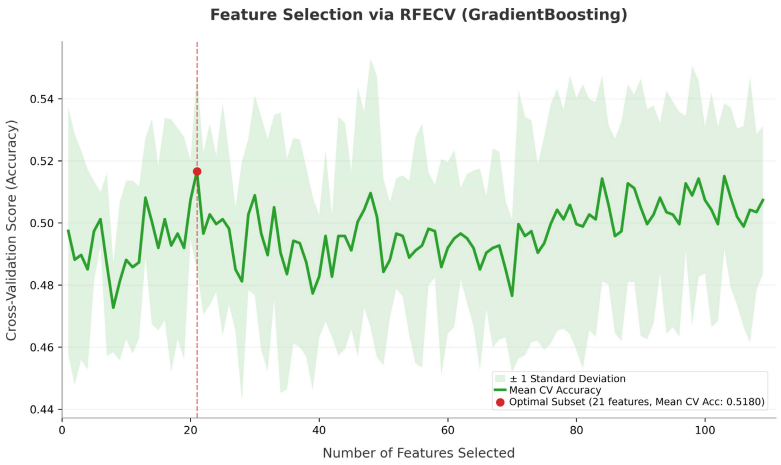


Fig. 2. Feature Selection via RFECV for the Gradient Boosting model. The green line indicates the mean accuracy and the shaded area represents ±1 standard deviation. The red dot highlights the optimal subset of 21 variables. (Color figure online)

Table 2. Genetic Algorithm Configuration, Search Space, and Optimized Parameters for the Gradient Boosting Classifier.

Category	Parameter	Search Space/Value	Optimized Value
GA Settings	Population Size	25	–
	Max Generations	50	–
	Crossover Prob. (*cxpb*)	0.60	–
	Mutation Prob. (*mutpb*)	0.30	–
	Early Stopping Patience	25 generations	–
Search Space (Genes)	`n_estimators`	[100, 150, 200, 250, 300, 350]	100
	`learning_rate`	[0.01, 0.05, 0.08, 0.1]	0.01
	`max_depth`	[2, 3, 4, 5]	2
	`subsample`	[0.6, 0.7, 0.8, 0.9]	0.7
	`min_samples_split`	[5, 10, 15, 20]	20

Comparative Analysis: Exhaustive Grid Search vs. Evolutionary Optimization. To empirically validate the necessity of the penalized evolutionary approach, we benchmarked the Genetic Algorithm against a standard exhaustive Grid Search. The Grid Search evaluated the entire parameter space (1,536 combinations) utilizing a standard TimeSeriesSplit cross-validation, aiming to maximize the ROC-AUC score without any penalty for the generalization gap.

Table 3 presents the stark contrast between the two optimization paradigms.

The exhaustive Grid Search evaluated all 1,536 possible parameter combinations and converged to a highly aggressive configuration (`learning_rate`= 0.1, `max_depth` = 3). This resulted in an exceptional in-sample performance (Train

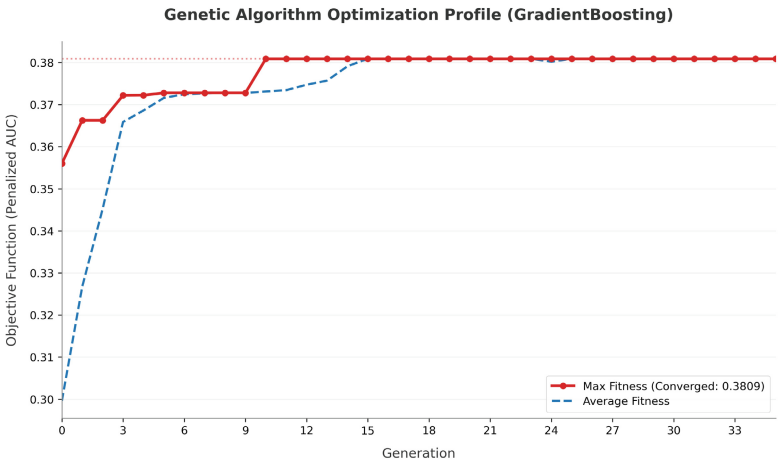


Fig. 3. Genetic Algorithm evolution profile. Absolute convergence occurred before the 40th generation, with the population average line (dashed) matching the global maximum of the penalized objective function.

Table 3. Performance comparison between Exhaustive Grid Search and the Evolutionary Genetic Algorithm for the Gradient Boosting classifier.

Metric/Parameter	Grid Search (Exhaustive)	Genetic Algorithm
Optimized Hyperparameters		
`n_estimators`	100	100
`learning_rate`	0.10	0.01
`max_depth`	3	2
`subsample`	0.6	0.7
`min_samples_split`	20	20
Statistical Performance		
Train AUC (In-Sample)	**0.9117**	0.7118
Validation AUC (In-Sample)	**0.9088**	0.5219
Test AUC (Out-of-Sample)	0.4716	**0.5041**

AUC > 0.91), creating a severe illusion of predictability. However, when exposed to the out-of-sample test set, the model's Test AUC dropped to 0.4716, falling below random guess level. This points to a known problem when applying traditional hyperparameter tuning methods to financial time series: the optimization algorithm ends up memorizing market noise. The Genetic Algorithm, guided by the gap-penalized fitness function (Eq. 6), leaned toward a more conservative setup (`learning_rate`=0.01, `max_depth`= 2). Instead of covering the full search space, the GA reached convergence through early stopping at generation 32. By accepting a lower in-sample accuracy (Train AUC 0.7118), the GA kept variance

under control and avoided picking up stochastic noise. As a result, the model was able to generalize to unseen test data, reaching a Test AUC of 0.5041 and keeping a positive predictive edge over the random walk baseline.

5.3 Financial Backtest and the Statistical-Financial Paradox

Accuracy and AUC-ROC measure statistical discrimination but fail to quantify the economic viability of a model, as they implicitly assume symmetric misclassification costs. To address this, we executed an out-of-sample financial backtest evaluating cumulative returns over the test set, contrasting the optimized Gradient Boosting with an identically evolved Multi-Layer Perceptron (MLP) neural network.

A critical mechanism implemented in our backtesting framework is the *Dynamic Threshold*. Financial sentiment probabilities rarely oscillate symmetrically around the theoretical 0.50 boundary. To prevent the strategy from being systematically biased toward long or short positions due to the inherent class probability shift of the evolved models, the decision threshold was dynamically calibrated as the median probability of the validation set predictions. For the Gradient Boosting model, the dynamic threshold was set at 0.5113, and for the MLP, at 0.5187.

Figures 4 and 5 illustrate the equity curves of the proposed strategies benchmarked against a passive Buy & Hold approach over the 279-day test period.

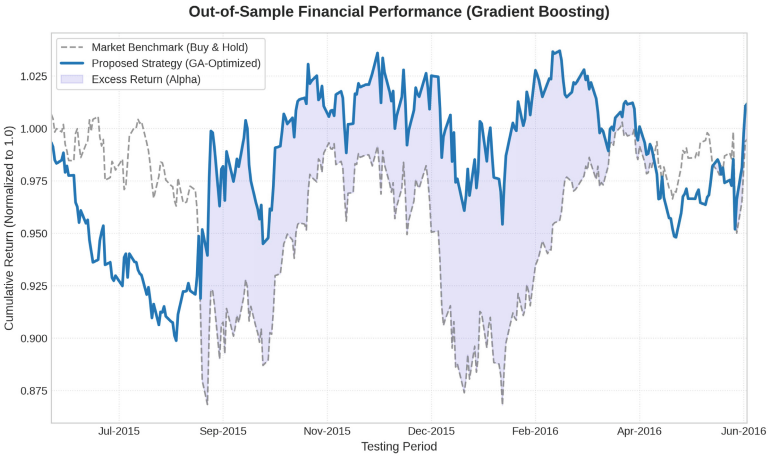


Fig. 4. Out-of-Sample equity curve for the GA-Optimized Gradient Boosting strategy. The model exhibited a highly conservative trading pattern, resulting in a marginal positive Alpha.

The Gradient Boosting model (Test AUC = 0.5041) exhibited a highly conservative behavioral profile. While it successfully navigated the market downturns

without catastrophic losses, its strict variance control prevented the algorithm from fully exploiting high-magnitude volatility events. Operating 189 d on the long side and 90 d on the short side, the strategy yielded a final return of 1.18%, translating to a modest, albeit positive, Alpha of **1.69 p.p.** against the market baseline (–0.51%).

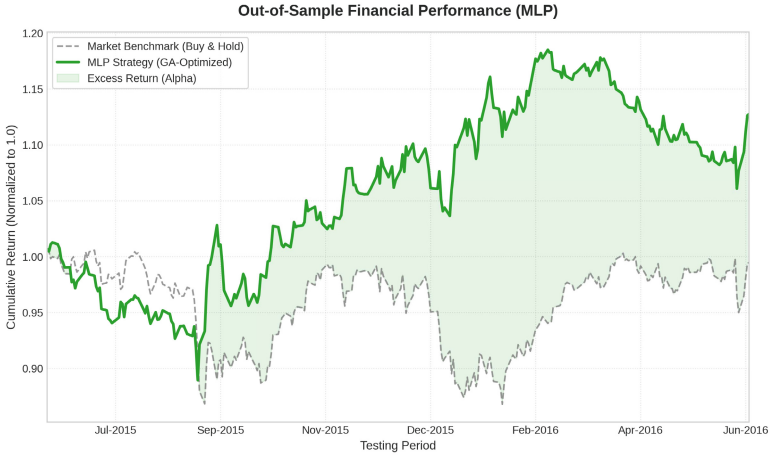


Fig. 5. Out-of-Sample equity curve for the GA-Optimized Multi-Layer Perceptron. Despite poor statistical performance, the neural network effectively captured high-magnitude asymmetric market shifts.

Conversely, the MLP presented a statistical metric indicative of failure (Test AUC = 0.4812). However, its financial simulation told a different story. Through evolutionary optimization, the GA identified a topology of 64–32 units, utilizing the `sgd` solver with a learning rate of 0.001 and an L_2 penalty (α) of 0.0001. Unbound by the rigid splits of decision trees, this architecture was able to capture non-linearities and asymmetric market anomalies dictated by the *FinBERT* sentiment signals. By identifying and trading on "fat-tail" events, the MLP generated a final return of 12.76%, reaching a $+$**13.27 p.p.** Alpha.

These empirical results highlight a critical challenge in computational finance: symmetric statistical metrics, such as AUC-ROC, can be insufficient on their own to attest to the economic viability of trading algorithms. As demonstrated by our findings, the practical success of NLP-driven models relies not necessarily on forecasting the majority of low-volatility days, but rather on capturing market asymmetries and utilizing dynamic probability thresholds to exploit extreme stochastic divergences.

6 Conclusions

This study tackled the problem of predicting the DJIA under noisy market conditions, where symmetric evaluation metrics fall short. To address this, we proposed LEAP, a leakage-free pipeline that brings together FinBERT sentiment, technical indicators, and temporal lags. We reduced overfitting through RFECV and a Genetic Algorithm guided by a generalization-penalized fitness function. Our results brought up an interesting statistical-financial paradox: while Gradient Boosting showed better discrimination (Test AUC 0.5041, Alpha +1.69 p.p.), the Multi-Layer Perceptron (MLP) statistically "failed" (Test AUC 0.4812) yet took advantage of market asymmetries via cost-sensitive learning to produce a **+13.27 p.p.** Alpha. These findings suggest that relying solely on symmetric metrics (e.g., AUC-ROC) can mislead the evaluation of trading algorithms, pointing to the need for magnitude-aware training. Future work could move toward a continuous Cost-Sensitive Learning framework, scaling instance weights by their volatility Z-score ($w_t = 1 + \alpha \frac{|R_t| - \mu_{|R|}}{\sigma_{|R|}}$). Bringing in intra-day data and sequence-to-sequence architectures may help improve alpha extraction from high-frequency news flows. Additionally, ablation studies isolating the individual contribution of each pipeline component, such as the sentiment features, the cost-sensitive weights, and the evolutionary optimization, would further strengthen the understanding of the framework and are left as a direction for future work.

Acknowledgment. The authors would like to thank FAPERGS (24/2551-0001396-2, 23/2551-0000773-8), CNPq (307416/2025-9), FAPERGS/CNPq (23/2551-0000126-8) and the Fesurv-University of Rio Verde.

References

1. Allen, F., Karjalainen, R.: Using genetic algorithms to find technical trading rules. J. Financ. Econ. **51**(2), 245–271 (1999)
2. Bailey, D.H., Borwein, J.M., Lopez de Prado, M., Zhu, Q.J.: Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting. Notices AMS **61**(5), 458–471 (2014)
3. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
4. Bragança, H., Pinto, R., Dalmazo, B., Lucca, G., Borges, E., Mattos, V.L.D., Berri, R.: Enhancing stock market predictions: the role of feature selection techniques in financial modeling. In: Gervasi, O., et al. (eds.) ICCSA 2025. LNCS, vol. 15649. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-96997-3_24
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Cerqueira, V., Torgo, L., Mozetič, I.: Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach. Learn. **109**(11), 1997–2028 (2020)
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. Cont, R.: Empirical properties of asset returns: stylized facts and statistical issues. Quantitative Finance **1**(2), 223–236 (2001)
9. Fama, E.F.: Efficient capital markets: a review of theory and empirical work. J. Financ. **25**(2), 383–417 (1970)
10. Gu, S., Kelly, B., Xiu, D.: Empirical asset pricing via machine learning. J. Financ. **75**(1), 2223–2273 (2020)
11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
12. Jegadeesh, N.: Evidence of predictable behavior of security returns. J. Financ. **45**(3), 881–898 (1990)
13. Lopez de Prado, M.: Advances in Financial Machine Learning. John Wiley & Sons (2018)
14. Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M.: Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. Appl. Soft Comput. **90**, 106181 (2020)
15. Souza, A.V.Set al.: Predictive analysis with technical indicators and features selection for futures contracts trading. In: Gervasi, O., et al. (eds.) ICCSA 2025. LNCS, vol. 15649. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-96997-3_22
16. Yang, Y., Uy, M.C.S., Huang, A.: Finbert: a large language model for extracting information from financial text. Contemp. Account. Res. **37**(2), 900–922 (2020)

A High-Throughput Real-Time Object Counting Framework for Industrial Conveyors: Asynchronous Pipelining and Spatiotemporal Consistency Strategy

Hieu Trung Nguyen, Thuan Huy Chau, and Vinh Dinh Nguyen(✉)

Department of Information Technology, FPT University, Can Tho Campus, Can Tho, Vietnam
{hieunt239,vinhnd29}@fe.edu.vn, thuanchce170133@fpt.edu.vn

Abstract. The automated counting of items on high-speed conveyor belts remains one of the critical activities in the logistics of Industry 4.0. Still, the implementation of automated systems in the Industrial IoT is greatly limited by several factors in the external environment of the systems such as dust, lighting, noise, and the significant delay caused by the traditional linear systems' sequential processing problem. This paper addresses the all-in-one problem of processing industrial object counting of all processing of data within the framework of high industrial throughput. We developed an Asynchronous Parallel Processing Architecture, wherein image capturing is separated from inference to fully utilize the processing powers of the GPU. The main detection engine is YOLOv11, equipped with the Spatiotemporal Consistency Framework, which is an object filtering post-processing unit that quickly eliminates temporary false positives by analyzing object death and trajectory without additional computationally heavy Re-ID features to track objects. We then defined an all-encompassing evaluation container of the proposed system that consists of Counting Accuracy (ACA), Mean Absolute Error (MAE), and RMSE. The test results of the framework achieved exceptional consistency with an average of 95.9% for the ACA. Furthermore, the proposed Asynchronous system demonstrated practical real-world throughput of 256 frames per second on the Tesla T4 GPU (Stress Test) and Edge-X RTX 2050 Laptop.

Keywords: Object Counting · YOLOv11 · Real-time Image Processing · Spatiotemporal Filtering · Industrial Automation · Ablation Study

1 Introduction

Product quantity tracking using conveyor belt technology is a problem whose accuracy is crucial for accurate inventory management and control of quality in production and distribution systems as they exist today [1, 7]. Manual tracking on conveyor belts is time consuming, laborious, and may cause human error [14]. Photoelectric sensors are efficient in general, given the simpler configuration of the load, but can only collect data partially due to colliding or overlapping objects which is common in cement bag

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 88–99, 2027.
https://doi.org/10.1007/978-3-032-30494-0_6

conveyors. Therefore, deep learning based object detection algorithms (i.e., YOLO family) have been the most popular alternatives for this problem because of their capability to detect objects rapidly with their single-stage architecture [2]. Despite the advantage of the YOLO family over conventional detectors thanks to speed itself, the use of these detectors in a factory environment introduces two significant persistent problems within the deployment stage, related to their detector: (1) Environmental noise effect on robustness: Factory cameras may be affected by motion blur, dust, etc., sudden illumination changes when obtaining photographs resulting in false positives in the detector, when nothing should be detected [9]. (2) Limitations imposed by edge device resource constraints: Typically the edge devices used within factories have limited computing resources. Complex models increase computational load and cause delays and result in dropped frames and lost counts in fast moving conveyor belts [15]. Therefore, in order to solve the problem over the system level, rather than solving the aforementioned problem through designing a new detector for it; we will solve both above issues. After deciding the number of model options used for our research, the YOLOv11n model [13] has a small size, and we designed an asynchronous pipeline around it to avoid I/O bottlenecking. Moreover, we applied a space-time consistency filter for checking all the detections against all existing frames to identify if an object persists in more frames and moves reasonably. It will not be allowed into the final count to any extent if it fails to match those standards, so it would lose selection.

2 Related Work

Consequently, beyond the previously mentioned counting methods that can be utilized at industrial scale three specific systems were used to address the problem.

Sensors with physical techniques, low-cost photoelectric or laser sensors are accurate but not when items are close to one another. This is a typical issue which you will see in high-speed production line, for example, when two bags of cement come into contact with one another their individual triggers will result in a single signal.

Tracking based counting, the initial generation of AI detectors to get started uses double data with both YOLOv5 and DeepSORT systems to track on the ground. DeepSORT was developed to employ extraction of appearance feature vectors (re-identification) for tracking. While this served well in many cases, it was ineffective at monitoring thousands of identical-looking packages (i.e. the Re-ID feature vector consumes high computational resources). For example, in Tariq [16], it was shown that if the training dataset was augmented by additional images generated synthetically from real factory floor data, YOLOv5 and YOLOv7 could be used to count industrial components; in contrast, Kumar et al. [17] enhanced the speed of counting with YOLOv8 but accuracy declined during the run as factory lighting changed part way through the run for this reason. We addressed this by running HSV jitter during the training process. Tracking-wise, [18] noted that it was reasonable to simply use positional overlap and not Re-ID feature vectors for association among the objects in a variety of frames of a fast moving conveyor belt and that it is this fact that motivated our use of ByteTrack. Furthermore, Jiang et al. [19] compared different YOLO versions and found that YOLOv11 was more efficient per-parameter than YOLOv8, which played an important role in our choice

in choosing YOLOv11 for the edge applications. Additionally, Liu et al. [20] reported YOLO-based detectors are non-reliable when operating in high vibration, dusty environments unless some form of post-processing is applied to them—a similar issue that the Spatiotemporal Consistency Filter we designed was to mitigate.

Therefore, this research introduces a solution that could be addressed using ByteTrack [5], which is a detector which does not use features of the object in view. It uses IoU (intersection over union) instead to associate objects - which significantly reduces the time needed to do the problem. Coupled with this method, a simpler and efficient solution to the noise characteristics of industrial feeds as compared with previous approaches is provided by Spatiotemporal Filter [10].

3 Methodology

3.1 System Architecture

The system has two separate subsystems that work independently: Frame Capture and Processing as shown in Fig. 1.

The Thread Acquire reads an image from the Camera Stream and loads it to Shared Buffer. Since we want to minimize Latency as much as possible, the FFmpeg Backend was setup with buffer_size = 0 (no buffering) and fflags = nobuffer (this prevents build up of frames on the backend so the decoder discards old frames). Therefore, only the latest frames are saved for additional usage. All frames are stored in a thread-safe FIFO Buffer (shared memory).

A second thread is in charge of processing. This thread gets the latest frames from the queue and performs all operations of the full Inference Pipeline. A Mutex will be used to provide a lock on the shared buffer in order to avoid contention between threads. Stress testing showed that without Synchronization, occasionally there were corrupted frames when running at 256 FPS.

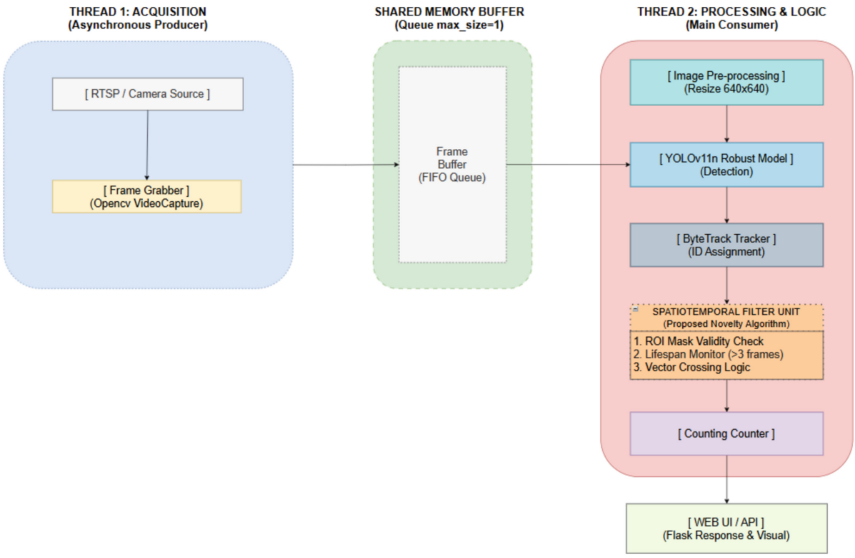


Fig. 1. Proposed system architecture diagram with an asynchronous pipeline.

3.2 Robust Detection Model (YOLOv11)

We selected YOLOv11n [13] based on its parameter efficiency. Training was tuned specifically for the conveyor environment rather than using default augmentation settings:

- **Photometric Distortions:** The HSV jittering $(h = 0.015, s = 0.7, v = 0.4)$ was applied to mimic sensor noise and different lighting and dust conditions in a factory.
- **Geometric Transformations:** For translating and scaling the objects (0.1, 0.2) without changing the orientation (which is a fixed parameter as objects are aligned) of the objects, we used horizontal flipping ($p = 0.5$) along with rotation being set to 0.0 degrees as it would not match the orientation the conveyor was set to, thus controlling the minor positional shifts on the belt without altering the objects.
- **Geometry-Preserving Strategy:** The value for Mixup and Mosaic was set to 0.0 as to not alter the natural edge boundaries of the neighboring objects within the real dataset [11] and thus not allowing the model to learn the features of the merged objects.

3.3 Spatiotemporal Consistency Framework

The raw bounding boxes from the yielded detector are inputted into the ByteTrack algorithm [5] to receive unique ID numbers. However, tracking, on its own, cannot perform accurate counting in noisy environments. We propose a specialized noise filter, consisting of two logical gates:

A. **Temporal Lifespan Filter**

In order to mitigate flickering False Positives caused by noise, we define the "Lifespan" (L_{id}) of a tracked object. An object can be seen as a candidate only if it sustains

for a minimum threshold of τ:

$$L_{id}^{(t)} = \begin{cases} L_{id}^{(t-1)} + 1, & \textit{if detected} \\ 0, & \textit{if lost} \end{cases}$$

We established $\tau = 3$ frames experimentally. Objects with $L_{id} \leq 3$ are discarded as transient noise. The reason for setting $\tau = 3$ is that since the belt moves at about one and half meters per second, and a frame from the camera is taken every thirty seconds, then any object will be in view for at least thirty frames. Therefore, it seems reasonable to set τ, or time constant for smoothing, as three because this would eliminate all of the transients which occur on fewer than two frames.

B. **Vector-based Crossing Logic**

To avoid double-counting of objects, we track an object's centroid as it moves from one video frame to another. a count will be incremented only if that centroid passes the virtual counting line in the correct direction—i.e. forward counts; crossing backward such as during a belt-reversal will be ignored:

$$C_y^{(t-1)} < Y_{line} \leq C_y^{(t)}$$

C. **Trajectory Smoothing**

YOLO bounding box coordinates can be noisy (jittery) on a per-pixel basis from one video frame to the next which will make it difficult to track the centroids close to the counting line. to reduce these noise levels while minimizing the impact of temporal delay, we have used exponential moving average (EMA) to smooth the detection coordinate:

$$P_t = \alpha.P_{raw} + (1 - \alpha).P_{t-1}$$

where P_{raw} is the current detection coordinate, P_{t-1} is the previous detection coordinate, and α is the smoothing factor set at 0.6.

D. **Evaluation Metrics**

Object Detection is a common task for which you are graded on your performance using Mean Average Precision (mAP). Counting can be used as an alternative metric to mAP but will require additional effort when counting error needs to be quantified [10]. We employ the following three counting metrics:

- **Mean Absolute Error (MAE):** This is simply the average counting error value.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |C_{pred}^{(i)} - C_{gt}^{(i)}|$$

- **Root Mean Square Error (RMSE):** RMSE penalizes larger differences from expected results.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{pred}^{(i)} - C_{gt}^{(i)})^2}$$

- **Average Counting Accuracy (ACA):**

$$ACA = 1 - \frac{\sum_{i=1}^{N}|C_{pred}^{(i)} - C_{gt}^{(i)}|}{\sum_{i=1}^{N}C_{gt}^{(i)}}$$

4 Experimental Results

4.1 Experimental Setup

These datasets include five videos of different states: there are 5000 frames for each sequence; medium condition for sequence B (moderate motion speeds/partial occlusions); difficult condition for sequence C (dust/vibrations from the camera/blur due to motion). These sequences have been manually annotated for correct true count. Performance in the best possible conditions was tested using a cloud based environment with an NVIDIA Tesla T4 GPU for training and evaluation studies. To test this method on a basic laptop, an NVIDIA GeForce RTX 2050 (4 GB) has also been tested to make sure if this is to apply in practice. It also allows testing of the possibility of using the solution for application in industrial edge computers. All experiments were performed using Python. Deep Learning Models were created by use of PyTorch. Object Detection utilized the Ultralytics YOLO framework. Object Tracking utilized ByteTrack [5]. The system was run in an Environment utilizing one or more GPUs for high throughput processing. Additional Libraries (OpenCV) were utilized to perform Image Preprocessing and Video Stream Handling. The total pipeline development and testing were done under a Standard Linux Based Environment to verify the Compatibility of Edge Deployment Scenarios.

4.2 Quantitative Analysis: Robustness & Counting Accuracy

Tables 1 and 2 compared our Proposed Framework against a Baseline method (YOLOv11 + Standard ByteTrack without Spatiotemporal Filtering).

Table 1. Counting Performance Comparison

Sequence	Ground Truth	Baseline Count	Proposed Count	Baseline Accuracy	Proposed Accuracy
Seq A (Easy)	205	205(0)	205(0)	100%	100%
Seq B (Medium)	205	170(−35)	180(−25)	82.9%	87.8%
Seq C (Hard)	205	205(0)	205(0)	100%	100%
Average	–	–	–	94.3%	95.9%

Analysis: The data illustrate a marked distinction between detection quality and count accuracy. It is worth noting that the Baseline model already had a 100% success rate for Seq C—which means that augmenting the training with HSV was enough to make it able to detect and count objects despite static dust being present. However, this was different when using Seq B—in which case rapid movements of the conveyor belt resulted in ByteTrack losing track of object Ids in mid-crossings. Our vector crossing recovery strategy successfully recovered ten of these lost counts. To summarize: while robust training can handle environmental noise as well as motion induced by tracking loss—both are covered by the spatiotemporal filter and their joint effect is the major cause of the failures we observed. Figure 2 show the training result of the proposed method over 50 epoches.

Table 2. Error Metrics (Overall)

Method	MAE ↓	RMSE ↓
BaselineBaseline	11.7	20.2
Proposed (Ours)	**8.3**	**14.4**

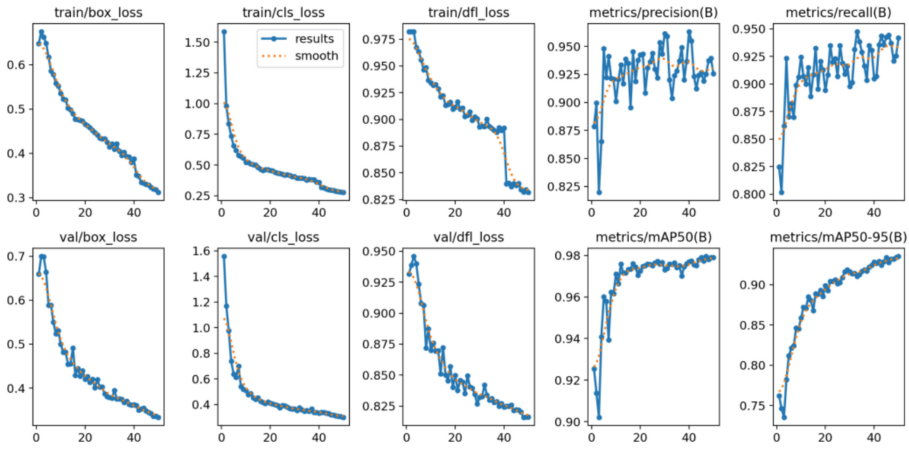


Fig. 2. Training performance graphs (mAP, Loss) of the YOLOv11n model over 50 epochs, demonstrating rapid convergence and stability.

4.3 Comparative Analysis: Efficiency vs. Accuracy

To justify the selection of YOLOv11 over its predecessor, YOLOv8 [4], we conducted a benchmark comparison as shown in Table 3.

Table 3. Efficiency Comparison (YOLOv8n vs. YOLOv11n)

Model	Parameters (M)	Complexity (GFLOPs)	mAP@50 (%)	Recall (%)	Precision (%)
YOLOv8n	3.01	8.1	97.9	93.0	93.9
YOLOv11n (Ours)	2.58	6.3	97.9	94.2	92.6

The results from Table 3 show that the mAP@50 scores for the two models were the same, each being 97.9%, which indicates that model performance had saturated in terms of feature learning for that specific problem. These two reasons support why we selected YOLOv11n for this work:

- Higher Recall: Higher recall than precision is needed in our proposed solution because while false positives may be rejected via back-end processing, false negatives (i.e., missed detections) cannot be recovered.
- Lower Computational Complexity: YOLOv11n is significantly less computationally intensive than its competitor, requiring 14.3% fewer parameters (approximately 2.58 million) versus approximately 3.01 million and utilizing a factor of 0.77 less GFLOPs (6.3 vs 8.2). This will allow our system to utilize more system resources on spatiotemporal filtering and video processing without experiencing bottlenecks with current computing architectures.

4.4 Component Analysis (Ablation Study)

To investigate the contribution of each module, we conducted an ablation study. We defined four configurations:

- **M1:** YOLOv11 Baseline (Standard Training, No Filter)..
- **M2:** M1 + Robust Training (Augmentation).
- **M3:** M2 + Temporal Lifespan Filter.
- **M4:** M3 + Vector Crossing Logic (Full System (Table 4)).

Table 4. Ablation Study Results

Config	Component	mAP@50	Average Counting Accuracy	FPS (Throughput)
M1	YOLOv11 Base	97.0%	85.0%	~278
M2	+ Robust Train	97.8%	94.3%	~278
M3	+ Lifespan Filter	97.8%	95.0%	~270
M4	**+ Vector Logic**	**97.8%**	**95.9%**	**~256**

Discussion: Robust Training (M2) gave the biggest single jump—85% to 94.3%—by fixing detection failures in dusty conditions. Adding the Spatiotemporal Filter (M3 → M4) pushed accuracy to 95.9%, mainly by cleaning up Seq B miscounts. The throughput cost was 22 FPS (278 → 256), corresponding to roughly 0.35 ms per frame—acceptable overhead given the accuracy gain.

4.5 Detailed Latency Analysis

Real-time performance is a key requirement. We analyzed the processing time breakdown per frame on a Tesla T4 GPU (Table 5).

Table 5. Latency Breakdown per Frame

Stage	Operation	Time (ms)	Percentage
1	Pre-processing	0.3 ms	6.4%
2	Inference (YOLOv11n)	3.9 ms	83.0%
3	Tracking & Logic	0.5 ms	10.6%
Total	End-to-End Latency	4.7 ms	100%

The Spatiotemporal Logic's proposed additional overhead is only 0.5 ms (as 10% of total pipeline time), and additional overhead is not relevant to counting accuracy which is most benefitted by further analyses. The overall throughput is $\approx$ 256 FPS (1000 ms/3.9 ms for inference limited throughput).

4.6 Comparative Analysis with State-of-the-Art Methods

To confirm the pros of the suggested structure, an analysis against two common industry counting configurations was performed: (1) YOLOv5 along with DeepSORT (which is a typical Re-ID based tracker), and (2) undifferentiated YOLOv8n with ByteTrack, but again, without our Spatiotemporal Consistency strategy. As discussed in Table 6, DeepSORT does robust tracking, but because of heavy Re-ID, which demands significant appearance feature extraction, the system becomes highly compute-bound and runs at a throughput level which is likely to be unsatisfactory for fast lines (<100 FPS on edge devices). On the contrary, with ByteTrack and the lightweight Spatiotemporal Filter, our framework takes advantage of geometry-based association. Thus, expensive Re-ID computations can be omitted. In addition, our implementation based on YOLOv11n is better than the baseline YOLOv8n in the sense that it consumes 14.3% fewer parameters (2.58M versus 3.01M), which translates to a better allocation of resources to the asynchronous buffering system. This improved system maintains a higher accuracy (95.9%) at real-time throughput, 256 FPS, which is a better fit for resource-constrained industrial edge nodes.

Table 6. Qualitative and Performance Comparison with Existing Methods

Method	Detector	Tracker Type	Re-ID Cost	Robustness to Dust	Est. Efficiency
Traditional DL	YOLOv5	DeepSORT	High	Medium	Low FPS
Recent SOTA	YOLOv8n	ByteTrack	None (IoU only)	Low	High FPS
Proposed	YOLOv11n	ByteTrack + Spatiotemporal Filter	None	High	Very High FPS

4.7 Detailed Latency Analysis

As a way to provide evidence for the functionality of the system, you can direct your attention to the performance of the system placed in a registered cement production plant during the usual production schedule (14:13:54, Sunday). The test used an IP camera, which was placed at a distance of 2.5 m from a belt conveyor, which was transporting 50 kg sacks of cement. The location was characterized by, and typical to industrial settings, dust in the air, motor vibrations from the conveyor belt, and control system lighting that changed in an unpredictable manner.

In the screenshot (Fig. 3), the green polygon marks the ROI and the red line is the counting threshold. The system held stable IDs on the cement sacks (e.g., ID:55, ID:56) despite the dust haze, and correctly ignored background movement that did not cross the line. Objects crossing the red line with valid vector trajectories and sufficient lifespan are

accurately counted, with the current cumulative count displayed in the top-left corner ("COUNT: 40"), confirming the system's robustness in a live production setting.

Fig. 3. Screenshot of real-time counting system interface at a cement factory. The image shows the system's ability to detect and count cement bags even under challenging conditions of low contrast and dust. (Green polygon: ROI; Red line: Counting threshold; Blue boxes: Detected objects) (Color figure online)

5 Conclusion

We have submitted a real-time system for counting objects in a challenging industrial setting and a cement factory at that. Using a lightweight YOLOv11n system, we developed new spatiotemporal consistency logic and coupled that with an asynchronous processing system to overcome the layered problems created by environmental noise (dust/vibration) and the strict latency requirements of rapid moving conveyor systems. We achieved counting accuracy (ACA) of 95.9% at over 256 frames of real-time counting in a millisecond on a Tesla T4 GPU. Also, an ablation study of the noise suppression system added 0.35ms of latency; however, noise counting in specific counted objects improved. In the near term, we want to deploy and further optimize the system on very low-cost edge systems, specifically the NVIDIA Jetson Orin Nano. Also, we want to assess the incorporation of lightweight attention to substantially increase the system performance in high-density occluded counting.

References

1. Li, Z., Wang, J., Wang, Y.: Real-time computer vision in smart manufacturing: a review. J. Real-Time Image Proc. **18**, 1–16 (2021)
2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)

3. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475 (2023)
4. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023). https://github.com/ultralytics/ultralytics
5. Zhang, Y., et al.: ByteTrack: multi-object tracking by associating every detection box. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13682, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20047-2_1
6. Wojke, N., Bewley, A., Paulus, D.: Simple online and real-time tracking with a deep association metric. In: IEEE International Conference on Image Processing (ICIP), pp. 3645–3649 (2017)
7. Ciuti, G., Dario, P.: Visual quality inspection systems: a systematic literature review. IEEE Access **8**, 20123–20157 (2020)
8. Sun, X., Wu, P., Hoi, S.C.: Face-to-face overlapping object counting on conveyor belts. IEEE Trans. Ind. Inf. **15**(3), 1545–1555 (2019)
9. Amrilla, Y., Heryadi, Y.: Cement sack counter application using convolutional neural network on conveyor belt. In: International Conference on Information Management and Technology (ICIMTech), pp. 1–6 (2021)
10. Gao, G., Liu, Y., Wang, M.: Counting objects by tracking: a robust method for industrial conveyor belts. IEEE Trans. Instrum. Meas. **69**(7), 4221–4231 (2020)
11. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
12. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
13. Ultralytics: YOLO11 documentation and performance metrics (2024). https://docs.ultralytics.com/models/yolo11
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. I–511 (2001)
15. Hsu, H.-M., Huang, T.-W., Wang, G., et al.: Multi-camera tracking of vehicles based on deep features re-identification. IEEE Trans. Intell. Transp. Syst. **22**(12), 7650–7660 (2021)
16. Tariq, R.: Counting objects in an industrial environments using YOLOv5 and YOLOv7, Padua Thesis Archive (2023)
17. Kumar, K., Kumar, P., Panda, S.: A real-time object counting and collecting device for industrial automation process using machine vision. IEEE Sens. J. **23**(12) (2023)
18. Faruk, M.J.H., et al.: A deep learning framework for object detection and tracking: CNN-based parcel tracking on conveyor belts. In: IEEE 6th International Conference on Sustainable Technologies for Industry 4.0 (STI) (2025)
19. Sapkota, R., Karkee, M.: Ultralytics YOLO evolution: an overview of YOLO26, YOLO11, YOLOv8, and YOLOv5 object detectors. arXiv preprint arXiv:2510.09653 (2025)
20. Liu, W., et al.: YOLO-STOD: an industrial conveyor belt tear detection model based on Yolov5 algorithm, Scientific Reports (2025)

An Integrated Congestion-Aware Framework for Task Allocation and Path Planning in Multi-Agent Pickup and Delivery

Dariush Ebrahimi(✉), Meet Sable, Qing Dai, Yash Ahuja, and Michael Ajibola

Department of Computer Science and Physics, Wilfrid Laurier University, 75 University Ave W, Waterloo, ON N2L 3C5, Canada
debrahimi@wlu.ca, {sabl3770,daix1890,ahuj8260,ajib9610}@mylaurier.ca

Abstract. Large-scale Multi-Agent Pickup and Delivery (MAPD) systems operate in dynamic environments where agents must continuously allocate tasks and plan collision-free paths under strict real-time constraints. As system scale increases, congestion and agent interference become primary bottlenecks, significantly degrading throughput and increasing idle time. This paper presents a scalable lifelong MAPD framework that tightly integrates congestion-aware task allocation with priority-based path planning. Tasks are assigned using heuristic cost estimates that capture both spatial distance and congestion tendencies, reducing the likelihood of bottlenecks before planning occurs. For motion coordination, we extend Priority Inheritance with Backtracking (PIBT) with congestion-aware guidance and deviation refinement, enabling efficient local conflict resolution without global replanning. Extensive experiments in a realistic warehouse benchmark demonstrate that the proposed approach substantially improves task throughput, reduces agent idle time, and remains stable under extreme congestion. The framework scales effectively to scenarios with up to 3,000 agents while consistently meeting real-time execution limits, highlighting its suitability for large-scale robotic logistics systems.

Keywords: Multi-Agent Path Finding (MAPF) · Multi-Agent Pickup and Delivery (MAPD) · Task Allocation · Priority Inheritance with Backtracking (PIBT) · Congestion-Aware Planning · Large-Scale Multi-Robot Systems

1 Introduction

Multi-Agent Path Finding (MAPF) concerns the computation of collision-free paths for multiple agents operating in a shared environment while satisfying spatial and temporal constraints. MAPF serves as a foundational problem for numerous real-world applications, including automated warehouse robotics [1], intelligent transportation systems [2,3], and large-scale simulation and digital

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 100–115, 2027.
https://doi.org/10.1007/978-3-032-30494-0_7

entertainment. Over the past decade, MAPF has received significant research attention, leading to the development of both optimal and suboptimal algorithms capable of handling increasingly complex scenarios.

The classical MAPF problem [4] assumes a static setting in which all agents are assigned fixed start and goal locations prior to planning. In this formulation, agents move synchronously on a graph and must avoid vertex and edge conflicts. While single-agent path planning can be solved efficiently using algorithms such as A*, extending these approaches to multi-agent systems leads to an exponential growth in the search space. To address this challenge, optimal solvers such as Conflict-Based Search (CBS) were introduced, decomposing the problem into a high-level conflict resolution process and low-level individual path searches [5,6]. CBS and its variants have demonstrated strong performance for moderately sized problems, but often struggle in highly congested or large-scale environments due to the increasing number of agent interactions.

Beyond classical MAPF, many real-world systems operate in dynamic and online settings where tasks appear over time, and agents must continuously replan. This gives rise to the Multi-Agent Pickup and Delivery (MAPD) problem, which extends MAPF by incorporating task allocation alongside path planning. Early online MAPF approaches such as COBRA [7], Token Passing (TP), and Token Passing with Time Swaps (TPTS) [8] emphasize efficiency and real-time execution, though they sacrifice completeness. Subsequent methods, including Multi-Label A* [9] and Task Conflict-Based Search (TCBS) [10], improve solution quality and completeness but remain limited in scalability. More recent hybrid approaches combining MAPF with combinatorial optimization techniques, such as TSP-based task sequencing [11], demonstrate promise but often rely on computationally expensive centralized planning.

In parallel, decoupled and priority-based methods have emerged as practical solutions for large-scale systems. Among them, Priority Inheritance with Backtracking (PIBT) offers an efficient and decentralized mechanism for resolving local conflicts by dynamically propagating priorities among agents [12]. PIBT has shown strong scalability properties and robustness in dense environments, making it particularly suitable for real-time MAPD scenarios. However, when applied without explicit consideration of task allocation and congestion patterns, its performance can degrade as agent density increases.

This paper builds upon congestion-aware PIBT by integrating it with a tailored heuristic-based task allocation strategy. The proposed framework explicitly accounts for congestion tendencies during task assignment, thereby reducing conflict frequency and improving overall system throughput. By jointly addressing task allocation and path planning within a unified framework, the approach achieves improved stability and efficiency in highly congested, large-scale MAPD environments.

The remainder of this paper is organized as follows. Section 2 formally defines the MAPD problem considered in this work. Section 3 presents the proposed algorithm and analyzes its computational properties. Section 4 evaluates the

approach through extensive experimental studies and benchmark comparisons. Finally, Sect. 5 concludes the paper and discusses directions for future research.

2 System Model and Problem Definition

We consider a MAPD problem in a warehouse-like environment modeled as an undirected graph $G = (V, E)$, where V denotes a finite set of discrete locations and E denotes feasible connections between locations. Time is discretized into uniform steps $t \in \mathbb{N}$. The system consists of a set of agents $A = \{a_1, \ldots, a_n\}$, with $|A| \leq |V|$, and a set of pickup-and-delivery tasks $X = \{x_1, \ldots, x_k\}$.

Each agent occupies at most one vertex at any time. A task represents a pickup-and-delivery operation that requires an agent to visit a pickup location and subsequently deliver to a designated goal location. The problem is to dynamically allocate tasks to agents and compute collision-free paths such that the system operates efficiently in a shared and potentially congested environment.

Let $\pi_i[t] \in V$ denote the location of agent a_i at discrete time step t, and let $p_i[t]$ denote the priority of agent a_i at time t. Priorities are used as a mechanism for resolving conflicts during path execution.

Each task $x_i \in X$ is defined by a pickup location $s_i \in V$ (optional for tasks that begin immediately) and a delivery location $g_i \in V$. The objective of the MAPD problem is to maximize the system *throughput*, defined as the number of tasks completed within a designated time horizon, while satisfying a set of feasibility constraints.

At any time, each agent may be assigned at most one task. Agent motion is constrained such that, at each time step, an agent may either move to an adjacent vertex or remain at its current location, i.e.,

$$\pi_i[t+1] \in \text{Neigh}(\pi_i[t]) \cup \{\pi_i[t]\},$$

where $\text{Neigh}(v)$ denotes the set of vertices adjacent to v.

Paths must be collision-free. Specifically, vertex conflicts are not allowed, meaning that no two agents may occupy the same vertex at the same time:

$$\forall i \neq j,\ \pi_i[t] \neq \pi_j[t].$$

In addition, swap conflicts are prohibited, where two agents exchange positions simultaneously:

$$\forall i \neq j,\ (\pi_i[t] = \pi_j[t+1]) \wedge (\pi_i[t+1] = \pi_j[t]).$$

Task execution is subject to precedence constraints: for any assigned task x_i, the corresponding agent must visit the pickup location s_i before reaching the delivery location g_i. Agent priorities $p_i[t]$ may evolve over time to support decentralized and dynamic conflict resolution during execution.

A solution to the MAPD problem consists of two tightly coupled components. First, a task allocation policy dynamically assigns available tasks to agents based

on the current system state and heuristic cost measures. Second, a path planning mechanism computes, for each agent a_i, a collision-free trajectory $\{\pi_i[t]\}_{t=0}^{T_i}$, where T_i denotes the time step at which the agent completes its assigned task by reaching the delivery location g_i.

The proposed framework jointly addresses task allocation and path planning in order to improve throughput and scalability in large-scale, congested MAPD environments.

3 Algorithm Design and Analysis

This section presents the proposed lifelong planning framework for large-scale MAPD, including conflict-aware path planning, the task allocation strategy, and an analysis of computational complexity. Algorithm 1 serves as the main control loop and determines the next action for each agent at every discrete time step.

3.1 Lifelong Planning Framework

The system operates in a lifelong planning setting, where agents continuously update their paths in response to newly assigned tasks, congestion, and conflicts. At each timestep, the planner determines the next vertex $\pi_i[t+1]$ for every agent a_i based on its current position, task status, and local traffic conditions.

Algorithm 1 begins with a preparation phase (lines 1–7). During this phase, heuristic tables storing shortest-path distances to goal vertices are initialized or updated when time permits (lines 2–4). Agents that are newly initialized or assigned a new task are marked for replanning (lines 5–7), ensuring that planning effort is focused only on agents that require updated guidance.

Guided path planning is performed in lines 9–14. For each marked agent, a new path is generated using the congestion-aware planner described in Algorithm 2. After a path is computed, congestion flow values are accumulated along the planned trajectory (line 13), representing anticipated occupancy by higher-priority agents. These flow values bias subsequent planning calls and reduce the likelihood of congestion buildup.

The refinement phase, shown in lines 16–21, addresses agents that deviate from their guided paths. Agents are sorted by descending deviation magnitude (line 16), where deviation is defined as the difference between the agent's current position and its guided path [12]. For each deviating agent, replanning is invoked again using Algorithm 2 (line 19), followed by flow updates (line 20). This refinement mechanism allows agents to gradually return to efficient trajectories without requiring global replanning.

Deviation handling is illustrated in Fig. 1. Each vertex is annotated with a pair $\langle d_{\text{dev}}, d_{\text{goal}} \rangle$, representing the deviation cost from the guided path and the remaining distance to the goal, respectively. During refinement, agents locally minimize this combined cost to rejoin their guided paths. Deviations may occur due to limited planning time, detours introduced by PIBT conflict resolution, or interference from higher-priority agents undergoing replanning.

Fig. 1. Deviation handling during refinement. Vertices are labeled with ⟨deviation, distance-to-goal⟩ costs guiding agents back to their planned paths [12].

Conflict resolution is handled in lines 21–25 using Priority Inheritance with Backtracking (PIBT). Agents are processed in descending order of priority (line 23). When a conflict is detected, priorities are dynamically updated to ensure progress for blocked high-priority agents.

Figure 2 illustrates the priority inheritance mechanism. When a high-priority agent is blocked by a lower-priority agent, the lower-priority agent temporarily inherits the higher priority and moves first to vacate the blocking vertex. This priority propagation may extend through a chain of agents and enables decentralized conflict resolution without global replanning, as implemented in lines 21–25 of Algorithm 1.

In highly congested environments, priority inheritance alone may be insufficient. Figure 3 depicts such a scenario, where no admissible forward move exists even after priority propagation. In these cases, backtracking is employed, causing one or more agents to retreat along their recent trajectories. This controlled backward motion releases congestion and restores feasibility. Backtracking is invoked as part of the PIBT process when forward movement is impossible, ensuring deadlock-free execution under extreme congestion [13].

All computationally expensive operations in Algorithm 1 are strictly bounded by a predefined time budget. This guarantees real-time responsiveness and fairness, which is essential for competitive benchmark environments such as the League of Robot Runners [14,15]. The algorithm outputs the next action $\pi_i[t+1]$ for each agent at line 26.

Algorithm 1. Lifelong Path Planning Framework

```
1: for a_i ∈ A do
2:    if t_cur < t_max then
3:       ht ← InitOrUpdate(ht, a_i)
4:    end if
5:    if status(a_i) ∈ {New, Reassigned} then
6:       M ← M ∪ {a_i}
7:    end if
8: end for
9: for a_i ∈ M do
10:   if t_cur < t_max then
11:      π_i ← PLAN(a_i)
12:      C ← UpdateCongestion(C, π_i)
13:   end if
14: end for
15: for a_i ∈ A sorted by δ_i ↓ do
16:   if t_cur < t_max then
17:      π_i ← PLAN(a_i)
18:      C ← UpdateCongestion(C, π_i)
19:   end if
20: end for
21: for a_i ∈ A sorted by p_i ↓ do
22:   if DetectConflict(π_i) then
23:      P ← PIBT(P)
24:   end if
25: end for
26: return {π_i[t + 1] | a_i ∈ A}
```

3.2 Congestion-Aware Path Planning

Algorithm 2 details the congestion-aware path planner used in the guided planning and refinement stages of Algorithm 1. The planner is a modified A* search that incorporates congestion penalties in addition to standard distance heuristics.

Lines 1–2 initialize the search at the agent's current location $v_{s_i} = \pi_i[t]$, setting the path cost g to zero and the heuristic value h using the precomputed table ht. The start node is inserted into the open list (line 3). The main search loop (lines 4–13) expands nodes until the goal vertex v_{g_i} is reached.

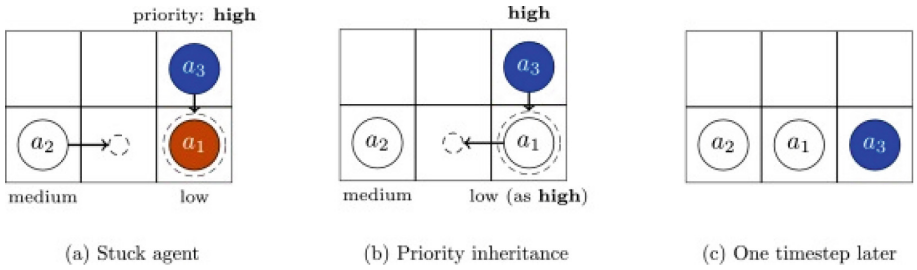
For each expanded node v_{curr} (line 5), all neighboring vertices v_{next} are examined (line 6). The evaluation value combines traversal cost, heuristic distance, and congestion penalties derived from traffic flow. If v_{next} is reached for the first time or a lower-cost path is found, node information and parent pointers are updated (lines 7–10), and the node is reinserted into the open list.

After reaching the goal, the planner extracts the least-congested path from the search tree (line 13) and returns it as π_i (line 14).

Algorithm 2. Congestion-Aware Path Planning (PLAN)

```
1: initialize start node at v_{s_i} = π_i[t] with g = 0, h = ht(v_{s_i})
2: push start node into open list
3: while goal v_{g_i} not reached do
4:    v_curr ← pop node with minimum evaluation value
5:    for each v_next ∈ Neigh(v_curr) do
6:       compute evaluation value using distance, heuristic, and flow penalties
7:       if v_next not visited or lower-cost path found then
8:          update node cost and parent
9:          push v_next into open list
10:      end if
11:   end for
12: end while
13: extract least-congested path π_i
14: return π_i
```

Traffic flow is defined as the number of higher-priority agents expected to traverse a vertex in a given direction [12]. Opposing flows are penalized more heavily than same-direction flows, as they are more likely to cause deadlocks.



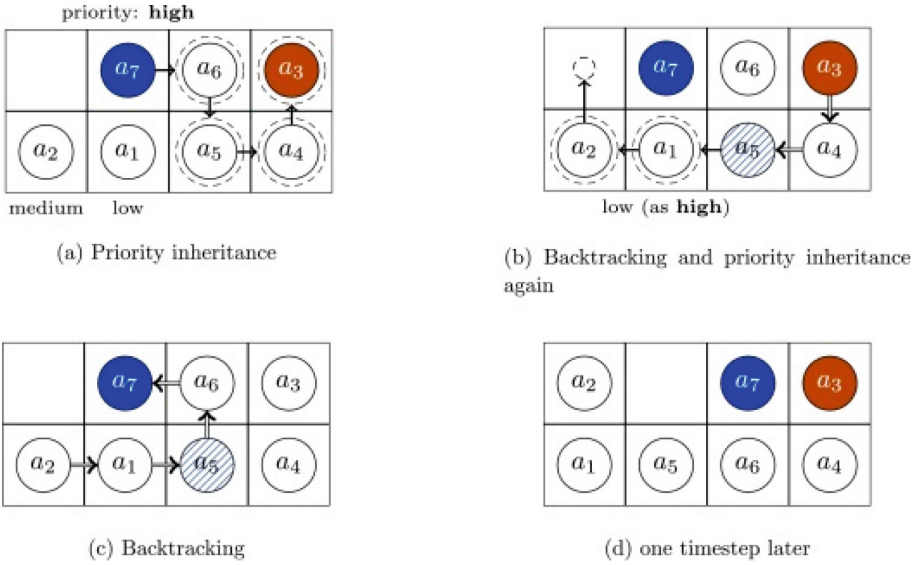
(a) Stuck agent (b) Priority inheritance (c) One timestep later

Fig. 2. Priority inheritance mechanism. A lower-priority agent temporarily inherits the priority of a blocked higher-priority agent to resolve conflicts locally [16].

3.3 Task Allocation

Task allocation is performed using the greedy heuristic-based procedure shown in Algorithm 3. At each planning cycle, the algorithm iterates over all vacant agents (line 1) and evaluates all unassigned tasks using a heuristic distance function (line 2). The task with the minimum heuristic value is selected and assigned to the agent (lines 3–5).

The heuristic computation uses BFS on the environment graph and caches results to avoid redundant searches. Once a task is assigned, the corresponding agent is immediately marked for replanning in the next iteration of Algorithm 1.



(a) Priority inheritance

(b) Backtracking and priority inheritance again

(c) Backtracking

(d) one timestep later

Fig. 3. Backtracking mechanism used when priority inheritance alone cannot resolve congestion. Agents retreat along recent trajectories to release space [16].

Algorithm 3. Heuristic-Based Task Allocation

```
for a_i ∈ A_vacant do
    x_j^* ← arg min_{x_j ∈ X_avail} ComputeHeuristic(a_i, x_j)
    if x_j^* ≠ NULL then
        task(a_i) ← x_j^*
        X_avail ← X_avail \ {x_j^*}
    end if
end for
```

3.4 Computational Complexity

Optimal MAPF is NP-complete, illustrated in Appendix A; Our proposed method prioritizes scalability and real-time feasibility via PIBT-based priority allocation. Each invocation of Algorithm repeated at a designated interval (e.g., 1 s). Algorithm 2 performs a modified A* search with worst-case complexity $O(|V| + |E| \log |V|)$. Since replanning is applied only to a subset of agents constrained by a time budget, the per-timestep complexity of the lifelong planner is approximately

$$O(|A| \cdot (|V| + |E| \log |V|)).$$

Task allocation (Algorithm 3) evaluates all unassigned tasks for each vacant agent. Each heuristic computation requires a BFS, yielding a worst-case complexity of

$$O(|A| \cdot |T| \cdot (|V| + |E|)).$$

All components run in polynomial time and are explicitly time-bounded, enabling the system to scale to thousands of agents while maintaining responsiveness in dynamic and congested MAPD environments.

4 Performance Evaluation

The MAPD problem has been widely studied, resulting in the development of numerous tools and benchmarks for systematic evaluation. To assess the performance of the proposed framework, we use the official simulation environment provided by the League of Robot Runners [15], a competitive benchmark designed to evaluate lifelong MAPD algorithms under realistic constraints. The simulator emphasizes robot dynamics, continuous task arrivals, online task allocation, congestion-aware planning, and strict real-time execution limits. This setting enables a robust evaluation of scalability and responsiveness in dynamic and highly congested warehouse-like environments.

4.1 Environment Setup

All experiments were conducted on a machine equipped with an Intel i7 9th-generation six-core CPU, 16 GB of RAM, and running Ubuntu Linux. The MAPF Start-Kit framework was used as the development and evaluation platform. This framework provides standardized interfaces for agent control, task allocation, and path planning, ensuring fair comparison across different MAPD approaches. Benchmark instances were generated using the MAPF Benchmark Archive, which supports flexible configuration of maps, agent counts, and task distributions.

Experiments were conducted on a large-scale sortation map of size 140×500, modeling a warehouse environment with narrow corridors and multiple pickup and drop-off locations. The map consists of blocked cells representing walls and obstacles, open cells that agents may traverse, designated pickup points, and delivery endpoints.

Tasks were specified using standard task files, where each task consists of a sequence of pickup and delivery locations that must be visited in order. Comment lines are ignored during parsing. For each experiment, three input files were generated: a map file defining the environment layout, a task file specifying pickup and delivery requests, and an agent file defining the initial positions of all agents.

To evaluate scalability, instances were generated with the number of agents ranging from 250 to 3,500 in increments of 250. Each simulation ran for 2,500 timesteps, with each timestep corresponding to one second of simulated time. At every timestep, the MAPD algorithm was allowed a maximum of one second of computation time to determine the next action for all agents.

Figure 4 shows a snapshot of the running simulation. Agents are represented as blue circles annotated with unique identifiers, and their facing directions are indicated by small dots. Open cells are shown in white, task destinations in

orange, and obstacles in black. New tasks arrive dynamically throughout the simulation, and vacant agents are assigned tasks online using the heuristic-based task allocation algorithm described in Sect. 3.

At each timestep, agents compute their next actions by integrating task assignment, congestion-aware path planning, deviation refinement, and priority-based conflict resolution using PIBT. This snapshot illustrates the dense interactions among agents navigating narrow corridors while executing tasks in real time.

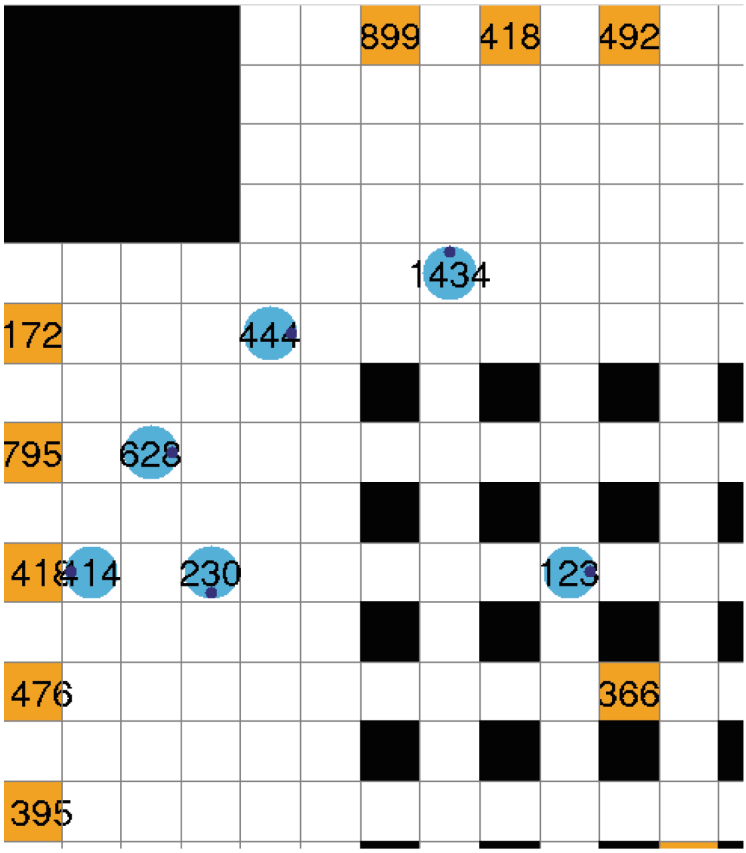


Fig. 4. Snapshot of the MAPD simulation on the sortation map. Agents navigate dynamically while executing pickup and delivery tasks under congestion and conflict constraints. Agents are represented as blue circles annotated with unique identifiers, and their facing directions are indicated by small dots. Open cells are shown in white, task destinations in orange, and obstacles in black. (Color figure online)

At the end of each simulation run, detailed output logs were generated. These logs include per-agent statistics such as average task completion time, number

of completed tasks, and idle time, as well as global metrics including the total number of tasks completed, average task waiting time, and computational performance. In particular, we measured the execution time of the MAPD algorithm at each timestep and the average computation time spent on task allocation and path planning.

The primary objectives of the evaluation were to assess scalability as the number of agents increases and to verify that the algorithm consistently operates within the one-second time constraint while maintaining high throughput.

4.2 Numerical Results

Figure 5 illustrates the relationship between the number of completed tasks and the number of agents over 2,500 timesteps. As expected, task throughput initially increases with the number of agents due to improved parallelism. However, the reference algorithm exhibits a noticeable performance drop at around 3,000 agents, while the proposed approach maintains a steady upward trend. This indicates improved robustness under high agent density.

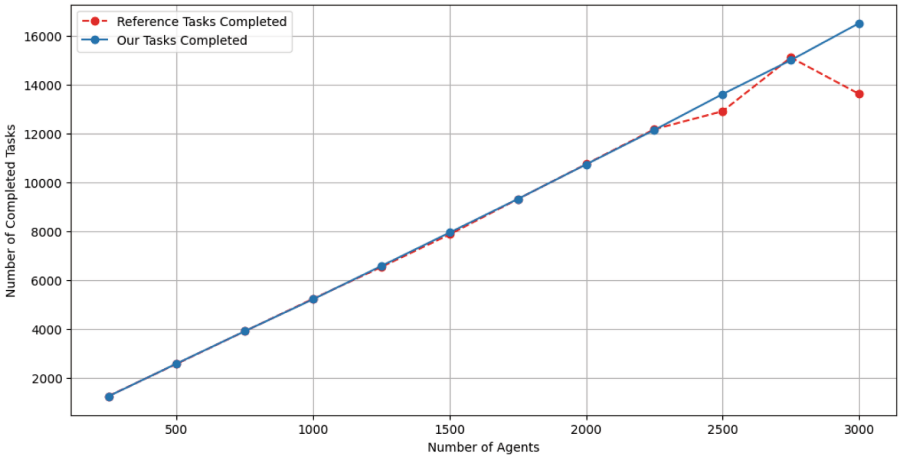


Fig. 5. Number of completed tasks versus number of agents over 2,500 timesteps (semilogarithmic scale).

Figure 6 plots the number of idle agents as the total agent count increases. The results show an exponential growth in idle agents beyond a certain scale, reflecting diminishing returns due to limited task availability and increased congestion. Nevertheless, the proposed algorithm mitigates this effect by more effectively distributing tasks and reducing congestion-induced blocking.

Figure 7 reports the average task completion time. While increasing the number of agents initially reduces completion time, a sharp increase is observed in

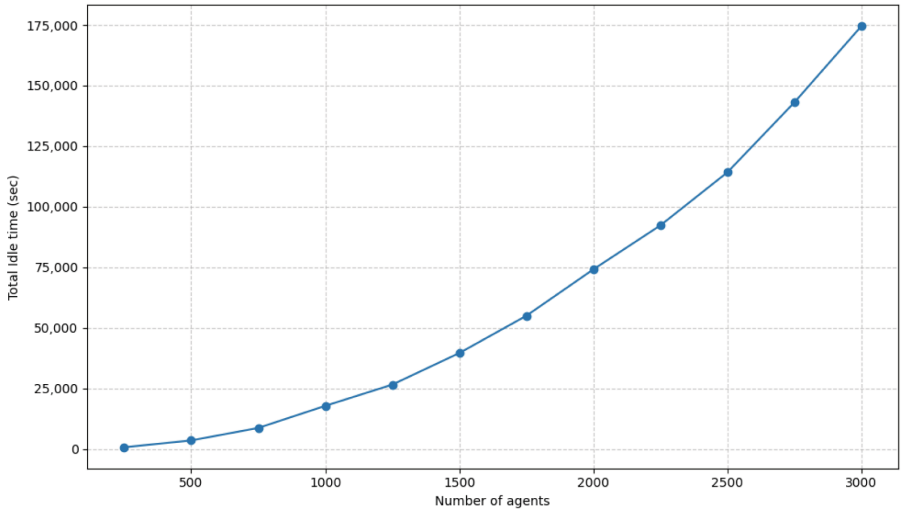


Fig. 6. Number of idle agents as a function of total agent count.

the reference algorithm at 3,000 agents, indicating severe congestion and inefficiencies. In contrast, our approach exhibits more stable behavior, suggesting improved scalability and congestion handling.

Due to memory limitations of the testing platform, simulations could not be executed beyond 3,000 agents. Nevertheless, the observed trends strongly suggest that the proposed method scales more gracefully than the baseline under extreme density.

5 Conclusion

This paper presents a scalable lifelong MAPD framework that integrates congestion-aware path planning, deviation handling, and heuristic task allocation with priority-based conflict resolution. Building upon the robustness of PIBT and its congestion-aware extensions, we introduce an enhanced task allocation strategy that significantly improves throughput under high agent density.

Experimental results demonstrate that the proposed approach consistently outperforms a strong reference implementation, successfully managing up to 3,000 agents on a large-scale sortation map of size 140×500. The improvements are particularly evident in highly congested scenarios, where task completion rates remain stable and average completion times degrade more gracefully.

Future work includes validating these results across additional benchmark maps and task distributions [14,17], as well as further investigating the causes of exponential growth in idle agents at large scales and how adaptive task generation or predictive allocation strategies may alleviate this limitation.

A Appendix: NP-Completeness of MAPF and MAPD

This appendix provides a concise proof sketch establishing the NP-completeness of the Multi-Agent Path Finding (MAPF) problem and its extension, the Multi-Agent Pickup and Delivery (MAPD) problem. The proofs follow standard complexity-theoretic arguments commonly adopted in the MAPF literature.

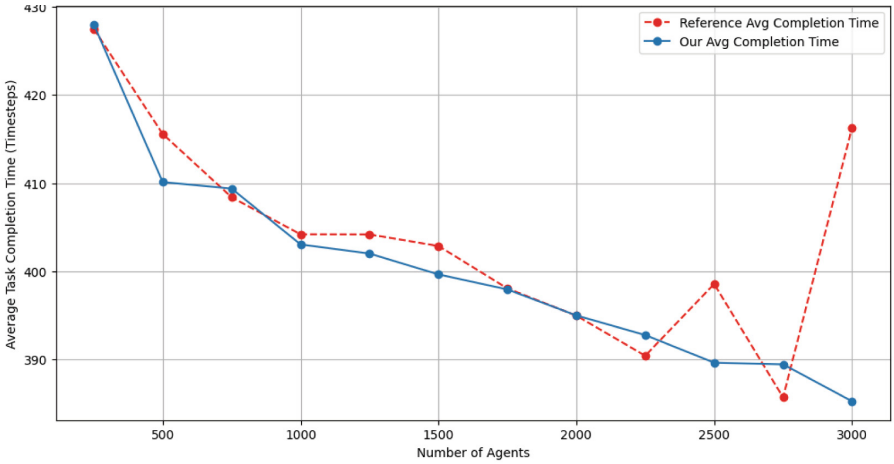


Fig. 7. Average task completion time as the number of agents increases.

A.1 NP-Completeness of MAPF

MAPF Is in NP. The MAPF problem belongs to the complexity class NP because any candidate solution can be verified in polynomial time. Given a set of paths, one for each agent, verification proceeds by checking two conditions. First, each path must connect the agent's start vertex to its goal vertex while respecting graph adjacency. Second, the paths must be collision-free, meaning that no two agents occupy the same vertex at the same timestep (vertex conflict) and no two agents traverse the same edge in opposite directions simultaneously (edge conflict).

Both checks can be performed efficiently. Path validity can be verified by scanning each agent's path, and collision detection can be performed by comparing agent positions at each timestep. Since the number of agents and timesteps is finite and bounded by the input size, the overall verification process runs in polynomial time.

MAPF is NP-Hard. MAPF is NP-hard via a polynomial-time reduction from the Sliding Tile Puzzle (also known as the 15-puzzle, Fig. 8), a well-known NP-complete problem.

The Sliding Tile Puzzle consists of an $n \times n$ grid populated by $n^2 - 1$ numbered tiles and a single empty cell. A move consists of sliding a tile into the empty space, and the objective is to transform an initial configuration into a specified goal configuration.

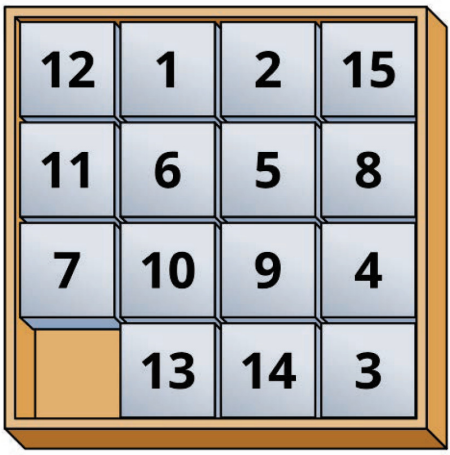


Fig. 8. Example of the 15-puzzle. The objective is to rearrange tiles into numerical order by sliding them into the empty space.

To reduce the Sliding Tile Puzzle to MAPF, we construct a MAPF instance as follows. Each grid cell corresponds to a vertex in the MAPF graph, and edges connect adjacent cells. Each numbered tile is treated as an agent whose start position corresponds to the tile's initial location and whose goal position corresponds to the tile's desired final location. The empty cell serves as the only free vertex, enforcing that agents can move only by swapping positions with this empty location, exactly mirroring the puzzle's movement constraints.

A collision-free solution to the MAPF instance corresponds directly to a valid sequence of tile movements solving the puzzle. Conversely, any solution to the puzzle induces collision-free paths for all agents in the MAPF instance. The construction involves a linear mapping of tiles and grid cells and can therefore be completed in polynomial time.

Conclusion for MAPF. Since the Sliding Tile Puzzle is NP-complete, the reduction to MAPF is polynomial, and MAPF is in NP, it follows that MAPF is NP-complete.

A.2 NP-Completeness of MAPD

Problem Definitions. The MAPF problem takes as input a graph and a set of agents, each with a fixed start and goal location, and asks for collision-free

paths for all agents. The MAPD problem generalizes MAPF by introducing tasks, where each task consists of a pickup location and a delivery location, and agents must execute these tasks without collisions, possibly under time or cost constraints.

Reduction from MAPF to MAPD. To prove the NP-hardness of MAPD, we reduce MAPF to MAPD in polynomial time. Given an arbitrary MAPF instance, we construct a corresponding MAPD instance by assigning each agent exactly one task. For each agent a_i in MAPF, we define a task with pickup location equal to the agent's start vertex and delivery location equal to its goal vertex.

This construction forces each MAPD agent to execute precisely the same movement required in the MAPF instance. Any collision-free solution to the MAPD instance, therefore, induces a valid solution to the original MAPF instance. Conversely, if the MAPF instance has no solution, then the corresponding MAPD instance is also infeasible.

The reduction is straightforward and linear in the number of agents, since it only introduces one task per agent and does not alter the underlying graph.

Conclusion for MAPD. MAPD is in NP because a candidate solution consisting of agent paths and task execution sequences can be verified in polynomial time. Since MAPF is NP-complete and reduces to MAPD in polynomial time, MAPD is NP-hard. Together, these results imply that **MAPD is NP-complete**.

References

1. Wurman, P.R., D'Andrea, R., Mountz, M.: Coordinating hundreds of cooperative, autonomous vehicles in warehouses. AI Mag. **29**(1), 9–19 (2008)
2. Li, J., et al.: Departure scheduling and taxiway path planning under uncertainty. In: AIAA Aviation 2019 Forum (2019)
3. Li, J., et al.: Scalable rail planning and replanning: winning the 2020 flatland challenge. In: Proceedings of the International Conference on Automated Planning and Scheduling, vol. 31, pp. 477–485 (2021)
4. Stern, R.: Multi-agent path finding - an overview. In: ARTIFICIAL INTELLIGENCE: 5th RAAI Summer School. Dolgoprudny, Russia (2019)
5. Wikipedia Contributors: Multi-agent pathfinding, Wikipedia, The Free Encyclopedia (2024). https://en.wikipedia.org/w/index.php?title=Multi-agent_pathfinding&oldid=1230323665
6. Kaduri, O., Boyarski, E., Stern, R.: Experimental evaluation of classical multi-agent path finding algorithms. In: AAAI Symposium on Combinatorial Search. https://ojs.aaai.org/index.php/SOCS/article/download/18560/18349/22079
7. Čáp, M., Vokřínek, J., Kleiner, A.: Complete decentralized method for on-line multi-robot trajectory planning in well-formed infrastructures. In: Proceedings of the International Conference on Automated Planning and Scheduling, pp. 324–332 (2015)

8. Ma, H., Li, J., Kumar, T., Koenig, S.: Lifelong multi-agent pathfinding for online pickup and delivery tasks. In: Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems, pp. 837–845 (2017)
9. Grenouilleau, F., van Hoeve, W.-J., Hooker, J.N.: A multi-label a* algorithm for multi-agent pathfinding. In: Proceedings of the International Conference on Automated Planning and Scheduling, pp. 181–185 (2019)
10. Henkel, C., Abbenseth, J., Toussaint, M.: An optimal algorithm to solve the combined task allocation and path finding problem. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4140–4146 (2019)
11. Kudo, F., Cai, K.: A TSP-based online algorithm for multi-task multi-agent pickup and delivery. IEEE Robot. Autom. Lett. **8**(9), 5910–5917 (2023)
12. Chen, Z., Harabor, D., Li, J., Stuckey, P.J.: Traffic flow optimisation for lifelong multi-agent path finding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 18 (2024)
13. Andreychuk, A., Yakovlev, K., Surynek, P., Atzmon, D., Stern, R.: Multi-agent pathfinding with continuous time. Artif. Intell. **305**, 103662 (2022)
14. Shen, B., Chen, Z., Cheema, M.A., Harabor, D.D., Stuckey, P.J.: Tracking progress in multi-agent path finding. arXiv preprint (2023)
15. Chan, S.-H., et al.: The league of robot runners: competition goals, designs, and implementation (2023). https://www.leagueofrobotrunners.org/
16. Okumura, K., Machida, M., Défago, X., Tamura, Y.: Priority inheritance with backtracking for iterative multi-agent path finding. Artif. Intell. **310**, 103752 (2022)
17. Stern, R., et al.: Multi-agent pathfinding: definitions, variants, and benchmarks. In: Proceedings of the International Symposium on Combinatorial Search, vol. 10, no. 1, pp. 151–158 (2021)

Bidirectional GRU Neural Network for Simultaneous Human Activity Recognition and Localization Using WiFi CSI Signals

Sakorn Mekruksavanich[1] and Anuchit Jitpattanakul[2(✉)]

[1] Department of Computer Engineering, School of Information and Communication Technology, University of Phayao, Phayao, Thailand
sakorn.me@up.ac.th

[2] Intelligent and Nonlinear Dynamic Innovations Research Center, Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
anuchit.j@sci.kmutnb.ac.th

Abstract. WiFi-based human sensing has emerged as a viable alternative for activity recognition and indoor localization, owing to its broad deployment, non-invasive operation, and inherent privacy preservation. This paper proposes a unified framework for concurrent human activity recognition and indoor positioning using WiFi channel state information (CSI) signals processed by bidirectional gated recurrent unit (BiGRU) networks. The framework is evaluated on the WiARL-UIT dataset—the first publicly released CSI dataset employing wireless timing synchronization—comprising 1,500 samples spanning six activity-location combinations. The pipeline covers data acquisition via software-defined radio (SDR) platforms, followed by CSI extraction, amplitude sanitization, and normalization. Classification is performed using BiGRU, which processes temporal sequences in both forward and backward directions. Stratified 5-fold cross-validation is used for performance assessment. The BiGRU model achieves 98.33% (±0.30%) accuracy, 98.33% (±0.30%) F1-score, and a loss of 0.0864 (±0.0383), surpassing all baseline architectures: CNN1D (90.80%), LSTM (90.00%), BiLSTM (95.67%), and unidirectional GRU (98.13%). The consistently low standard deviation across folds indicates stable, dependable generalization. Confusion matrix analysis further confirms strong discriminative ability across all activity-location classes with low misclassification rates. The results presented here suggest that BiGRU is particularly well-suited to the demands of WiFi CSI processing, given its capacity to capture temporal dependencies in both directions while remaining computationally efficient. Furthermore, the proposed framework affirms that simultaneous activity and location recognition is not merely a theoretical possibility but an achievable and practical objective—one that holds considerable promise for real-world deployment across domains such as smart home systems, elderly care monitoring, context-aware computing, and intelligent building management.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 116–128, 2027.
https://doi.org/10.1007/978-3-032-30494-0_8

Keywords: WiFi CSI · Human activity recognition · Indoor localization · Bidirectional GRU · Deep learning · Simultaneous recognition · Wireless sensing

1 Introduction

Human activity recognition (HAR) and indoor localization have emerged as significant research domains in ubiquitous computing, propelled by applications in smart homes, healthcare, eldercare, context-aware services, and intelligent building management [1]. Wearables, cameras, and dedicated infrastructure are traditional ways to sense things, but each has significant problems [3]. Wearables require users always to follow the rules, camera systems raise privacy concerns by capturing visual data, and dedicated solutions like infrared sensors or pressure mats are expensive to install and do not work well when more units are added.

WiFi-based sensing has become popular as a useful alternative, mostly because the necessary infrastructure is already in place in most indoor spaces [17]. As Wi-Fi signals move through a space, they bounce off objects and people, causing multipath fading. Channel state information (CSI) records these changes by taking highly precise amplitude and phase measurements across many subcarriers. This gives a much better picture of the channel than the received signal strength indicator (RSSI) does, and it also makes it easier to determine whether someone is there, moving, or doing something [2].

Even though WiFi-based activity recognition and localization have both made steady progress, not much work has been done on combining the two [12]. This is a significant gap because context-aware systems need to know both what a person is doing and where they are simultaneously. In smart home scenarios, distinguishing whether a user is seated in the living room or the bedroom enables appropriate environmental adjustments. In elderly care, simultaneous detection of a fall and its location accelerates emergency response. Nevertheless, most prior work addresses these tasks independently.

The core difficulty lies in extracting two distinct feature types from the same CSI measurements: activity-specific patterns that generalize across locations, and location-specific cues that remain stable across activities. Handcrafted feature methods have proven insufficient for this joint problem due to the complexity of manual feature design. Deep learning offers a more suitable path by autonomously learning hierarchical representations from raw signals. However, identifying a network architecture that effectively models CSI temporal dynamics while preserving computational efficiency remains an open challenge.

This paper proposes a unified framework for concurrent HAR and indoor positioning using WiFi CSI signals and bidirectional gated recurrent unit (BiGRU) networks. The framework is built upon the WiARL-UIT dataset, the first publicly available CSI dataset with wireless timing synchronization for joint recognition tasks. BiGRU processes input sequences in both forward and backward directions, allowing the network to exploit past and future context simultaneously when learning features that encode both activity and location information.

2 Related Works

This section surveys prior research in WiFi-based human sensing across three areas: activity recognition using WiFi signals, indoor positioning techniques, and the emerging joint task of simultaneous activity and location recognition. Relevant publicly available datasets are also discussed.

2.1 WiFi-Based HAR

WiFi-based activity recognition has drawn growing research interest due to its non-invasive operation and reliance on existing wireless infrastructure. Early studies demonstrated the feasibility of detecting human presence and basic motion using fluctuations in WiFi signal strength. More advanced methods use CSI to detect small multipath distortions that occur when a person moves, which helps recognize activities such as walking, sitting, and falling [15]. Later work used deep learning architectures, such as convolutional and recurrent networks, to automatically extract spatio-temporal features from CSI data to improve classification accuracy. Bidirectional long short-term memory (BiLSTM) networks, often used with convolutional neural networks (CNNs), have been especially good at modeling complex temporal dependencies in activity sequences [7]. Their ability to process sequences in both directions allows them to capture long-range dependencies, which are necessary for identifying subtle activity patterns [16]. A comparative analysis of BiLSTM and CNN with GRU models across various WiFi CSI datasets revealed that BiLSTM is superior in extracting long-term temporal structures from high-resolution data [15]. This temporal awareness enables reliable differentiation of activities, even when movement patterns are highly similar.

2.2 WiFi-Based Indoor Localization

A lot of research has been done on using WiFi to navigate indoors instead of GPS, which does not always work or can be wrong in enclosed spaces. Some methods use fingerprinting to compare observed signal patterns to a pre-made radio map, while others use trilateration and angle-of-arrival techniques based on geometric signal calculations. The use of CSI has greatly improved localization accuracy, enabling centimeter-level performance by leveraging fine-grained phase and amplitude data that RSSI-based methods cannot provide [4]. CSI gives a more detailed picture of the wireless channel than RSSI does. This makes sensing tasks like localization and activity recognition more stable and useful [7]. CSI also lets you see tiny changes in signal oscillations across subcarrier frequencies, which are important for accurate positioning and activity detection because they show changes in amplitude and phase [4]. Deep learning, especially convolutional networks, has improved localization by learning direct mappings from CSI features to spatial coordinates without requiring explicit signal-propagation models [9]. Recurrent architectures, especially LSTM networks, have also enabled modeling temporal dependencies in CSI streams. This makes localization more reliable in changing environments [8].

3 The HAR Framework

This section describes a unified framework for concurrent HAR and indoor positioning using WiFi CSI signals. As illustrated in Fig. 1, the framework comprises four sequential stages: data acquisition, preprocessing, classification, and evaluation. This structured pipeline enables systematic extraction and analysis of CSI patterns to predict human activities and their corresponding indoor locations jointly.

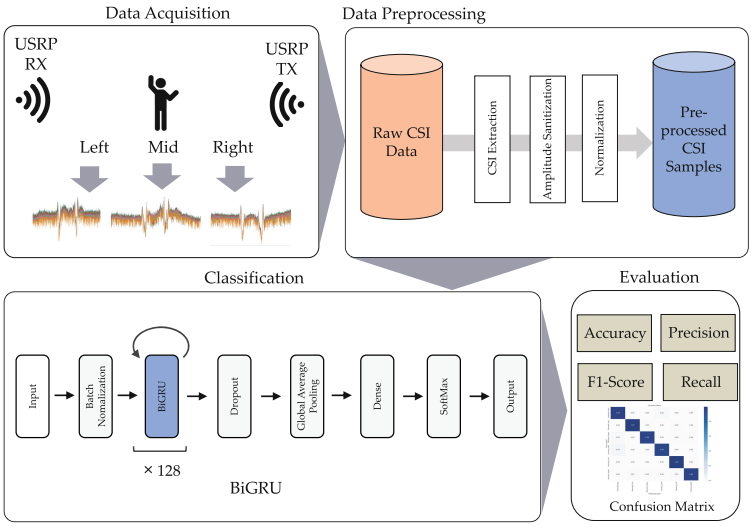


Fig. 1. The HAR framework based on CSI data used in this work.

The data acquisition stage employs a software-defined radio (SDR) platform with universal software radio peripheral (USRP) devices to capture raw WiFi CSI measurements from volunteers conducting various activities at different spatial positions. The pre-processing stage converts these raw measurements into clean, standardized time-series sequences via CSI extraction, amplitude normalization, and sanitization, preparing the data for model training.

The classification stage applies a BiGRU network to model temporal dependencies in sequential CSI data by processing input in both forward and backward directions. The evaluation stage then quantifies framework performance using accuracy, precision, recall, F1-score, and confusion matrix analysis under stratified 5-fold cross-validation.

The subsections below detail each component: the WiARL-UIT dataset, preprocessing procedures, the BiGRU architecture, and evaluation criteria.

3.1 WiARL-UIT Dataset

The WiARL-UIT dataset, introduced by Thanh et al. [13], is the first publicly available WiFi CSI dataset using wireless timing synchronization for joint activity recognition and indoor positioning. It fills a significant gap in the literature, as prior datasets were either withheld from public release or relied on wired synchronization, which constrained their real-world applicability. Data collection used an SDR framework built on two USRP B200-Mini devices equipped with VERT-2450 Wi-Fi antennas, running GNU Radio and implementing the IEEE 802.11 protocol.

Signal transmission and reception relied on orthogonal frequency division multiplexing (OFDM). Each OFDM frame included two synchronization words for timing and frequency offset correction, followed by a payload of BPSK pilot signals for CSI extraction. Each OFDM symbol contained 64 subcarriers: 4 pilot subcarriers, 49 data symbols, and 11 guard bands for interference suppression. Wireless timing synchronization was achieved through a dedicated algorithm enabling precise transmitter-receiver coordination without physical cable connections, improving practicality over cable-dependent approaches used in earlier work.

The dataset covers two activities—"HandUp" and "ToTheLeft"—performed at three indoor positions: "Right", "Mid", and "Left". This yields six class labels combining activity and location: HandUp-Right (A), HandUp-Mid (B), HandUp-Left (C), ToTheLeft-Right (D), ToTheLeft-Mid (E), and ToTheLeft-Left (F). Each class contains 250 samples, producing a balanced total of 1,500 CSI samples. Representative CSI samples are shown in Fig. 2.

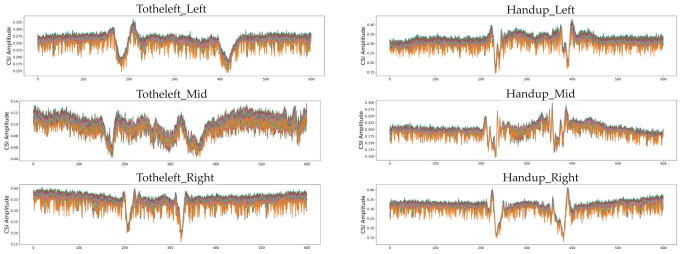


Fig. 2. Some samples of CSI data for the WiARL-UIT dataset.

The CSI measurements capture amplitude and phase variations in WiFi signals caused by human presence and motion. These variations result from multipath propagation, where signals reflect off the human body and surrounding objects, generating distinctive fingerprints for each activity-location combination. Accurate timestamping through wireless synchronization ensures proper alignment between transmitter and receiver, which is essential for temporal feature extraction. Together, the SDR-based collection setup and synchronization capability make the WiARL-UIT dataset a valuable benchmark for WiFi-based human sensing in smart environments and Internet of Things (IoT) applications.

3.2 Data Preprocessing

Preprocessing is an essential stage that converts raw WiFi CSI signals into a format suitable for deep learning training and evaluation. The raw CSI data from the WiARL-UIT dataset consists of complex-valued channel coefficients encoding amplitude and phase information across multiple subcarriers. If we do not handle these measurements correctly, noise, outliers, and inconsistencies can degrade the model's performance. A structured preprocessing pipeline is used to improve the signal and find useful features for classifying joint activity and location.

The first step is to get the CSI and break it down. The USRP device gets raw OFDM signals and uses GNU Radio demodulation blocks to get CSI from pilot subcarriers. Extraction is all about the pilots because each OFDM symbol has 64 subcarriers and 4 pilot positions. This helps you get the right channel estimates. The resulting coefficients are divided into two groups: phase and amplitude. Each part gives you a different piece of information about how the signal moves. Phase shows very small changes over time between channels, while amplitude shows how the signal strength changes when people are nearby and moving.

Next, amplitude sanitization is used to correct measurement errors and incorrect values. Hardware problems, interference, or packet loss can cause sudden changes in CSI amplitude data. We use a median filter with a 5-pixel-wide sliding window to smooth out sudden changes while preserving the signal's general shape. After that, a threshold-based method is used to remove outliers. This means that values more than 3 standard deviations away from the local mean are replaced with estimates from nearby subcarriers. This step ensures that the inputs are identical and that measurement noise has less effect during training.

Normalization is used to ensure that amplitude values are consistent across all samples and recording sessions. When the transmission power, receiver gain, or environment changes, the amplitude between sessions can vary significantly. Min-max normalization changes all the numbers so that they are all between 0 and 1, as shown below:

$$\mathrm{CSI}_{\mathrm{normalized}} = \frac{\mathrm{CSI} - \mathrm{CSI}_{\mathrm{min}}}{\mathrm{CSI}_{\mathrm{max}} - \mathrm{CSI}_{\mathrm{min}}} \tag{1}$$

where $\mathrm{CSI}_{\mathrm{min}}$ is the lowest amplitude value in each sample and $\mathrm{CSI}_{\mathrm{max}}$ is the highest amplitude value in each sample. Normalizing at the sample level keeps the relative structure of the amplitude signal intact and makes sure that all samples share a common scale. This is a trait that usually leads to more stable convergence during training.

3.3 BiGRU Network

Recurrent neural networks (RNNs) are a class of neural architectures well-suited to processing sequential data, making them a good fit for time-series signals like CSI measurements. RNNs, on the other hand, have an internal hidden state

that keeps track of temporal context across input sequences of different lengths [11,14].

The GRU, introduced by Cho et al. [5], is an improved version of the RNN designed to address the problem of gradient disappearance in regular RNNs. This is achieved through a gating mechanism that governs information flow across time steps. The GRU comprises two gates: an update gate, which controls how much the hidden state is refreshed, and a reset gate, which determines how much prior information is discarded. Together, these gates enable efficient modeling of long-range temporal dependencies in signals with complex sequential structure [6].

The BiGRU extends the standard GRU by processing input sequences in both forward and backward directions, allowing the network to incorporate both past and future context at each time step [10]. As shown in Fig. 3, a BiGRU layer consists of two parallel GRU hidden layers: one traversing the sequence left-to-right and the other right-to-left. Their outputs are concatenated to form the final layer representation.

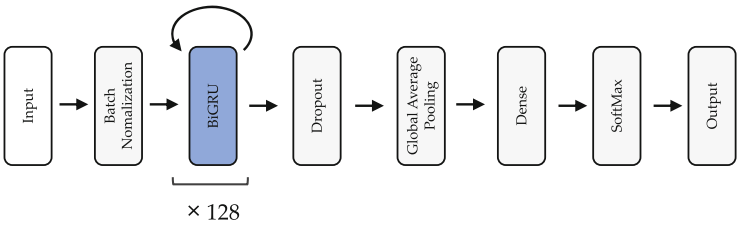


Fig. 3. Architecture of a BiGRU network used in this work.

For WiFi CSI-based joint recognition, BiGRU offers four key advantages. First, it captures both short- and long-term temporal dependencies in sequential signal patterns. Second, bidirectional processing allows the model to identify features that depend on both preceding and subsequent signal characteristics. Third, GRU's reduced parameter count relative to LSTM leads to faster training and lower overfitting risk, particularly valuable with limited data. Fourth, the bidirectional structure supports comprehensive automatic feature extraction, reducing reliance on manual feature engineering.

4 Experimental Results and Discussion

This section presents an evaluation of the proposed BiGRU-based framework for concurrent human activity recognition and indoor positioning using WiFi CSI signals. The experiments are designed to validate the BiGRU architecture and assess its advantage over alternative deep learning approaches in modeling temporal CSI patterns.

4.1 Experiments

All experiments were conducted on Google Colab Pro+, utilizing a Tesla V100 GPU for high-performance neural network training. Python 3.6.9 was used to build the software environment, and TensorFlow 2.2.0 served as the deep learning backend, with Keras 2.3.1 as the high-level modeling interface. Scikit-Learn was used for metric computation and validation; NumPy 1.18.5 for numerical and array operations; and Pandas 1.0.5 for data management and preprocessing. This stack provided a solid foundation for building and testing all the models.

To get an accurate estimate of performance, a stratified 5-fold cross-validation approach was used. The dataset was split into five equal parts, and the class and location distributions remained the same across all parts. In each iteration, four folds were utilized for training and one for testing. This process was performed five times, such that each fold was used as the test set once. The final performance measurements were reported as mean values with standard deviations, making it easy to see both overall performance and consistency across partitions.

We used the Adam optimizer with a starting learning rate of 1×10^{-3}. This made convergence faster and more stable than normal stochastic gradient descent. The batch size was set to 32 to balance training stability and computational cost. Models were trained for up to 200 epochs, with early stopping triggered when validation loss failed to improve for 15 consecutive epochs. The weights yielding the lowest validation loss were retained as the final model—categorical cross-entropy served as the loss function, appropriate for multi-class settings. Dropout regularization with a rate of 0.5 was applied to fully connected layers, randomly turning off half the neurons at each step to promote generalization and reduce overfitting.

To assess the BiGRU architecture's effectiveness, four baseline models were included for comparison: (1) CNN1D, a one-dimensional convolutional network extracting spatial features from CSI amplitude sequences; (2) LSTM, a standard recurrent network modeling temporal dependencies in sequential CSI data; (3) BiLSTM, a bidirectional LSTM processing sequences in both directions; and (4) GRU, a unidirectional gated recurrent network. All baselines were configured with comparable network depth and hidden unit counts to ensure a fair architectural comparison. Identical hyperparameters and cross-validation procedures were applied across all models to isolate the effect of architectural design on classification performance.

4.2 Experimental Results

The experimental findings confirm that the proposed BiGRU architecture outperforms all other tested models in jointly recognizing human activities and determining locations through WiFi CSI data. Table 1 summarizes the full set of evaluation metrics for every deep learning architecture examined. These metrics include accuracy, loss, and F1-score, along with their standard deviations computed across the 5-fold cross-validation.

Table 1. Performance metrics of deep learning models

Model	Performance		
	Accuracy	Loss	F1-score
CNN1D	90.80% (±8.37%)	0.58 (±0.57)	90.68% (±8.51%)
LSTM	90.00% (±11.19%)	0.28 (±0.23)	89.76% (±11.60%)
BiLSTM	95.67% (±1.30%)	0.18 (±0.05)	95.67% (±1.29%)
GRU	98.13% (±0.62%)	0.13 (±0.05)	98.13% (±0.61%)
BiGRU	98.33% (±0.30%)	0.09 (±0.04)	98.33% (±0.30%)

A noticeable trend emerges when comparing models of increasing complexity. Simpler designs yield lower performance, while more advanced recurrent architectures consistently achieve better results. The CNN1D architecture achieved 90.80% accuracy (±8.37%) but suffered a comparatively high loss of 0.58 (±0.57). This suggests that one-dimensional convolutional networks can derive meaningful spatial representations from CSI data. However, they fall short in modeling temporal variations, which are critical for reliable activity classification and position estimation. The conventional LSTM network produced a similar accuracy of 90.00% (±11.19%). Its higher variance across folds, however, points to less dependable training behavior when evaluated on different data subsets.

Introducing bidirectional processing into the LSTM framework led to a notable performance gain over the single-direction variant. The BiLSTM model attained 95.67% accuracy (±1.30%) alongside a considerably reduced loss of 0.18 (±0.05). This enhancement highlights the value of incorporating context from both forward and backward temporal directions when interpreting WiFi CSI patterns. In such data, information both preceding and following a given time step contributes to more precise classification outcomes. The smaller standard deviation further reflects more stable results across different validation splits.

The GRU architecture delivered strong results, achieving 98.13% accuracy (±0.62%) with a loss of 0.13 (±0.05). These figures surpass those of both the CNN1D and the standard LSTM models. This advantage likely stems from the GRU's streamlined gate structure. It effectively models extended temporal relationships while using fewer trainable parameters than the LSTM. This reduction in parameter count may help mitigate overfitting and improve the model's generalization.

Among all architectures tested, the proposed BiGRU model produced the highest overall performance. It achieved 98.33% accuracy (±0.30%), the lowest loss at 0.09 (±0.04), and an F1-score of 98.33% (±0.30%). The notably small standard deviation reflects highly stable and dependable outcomes across all data partitions. These results indicate that combining the GRU's efficient gate design with bidirectional sequence processing yields the most effective feature representation for concurrent activity recognition and localization from WiFi CSI signals.

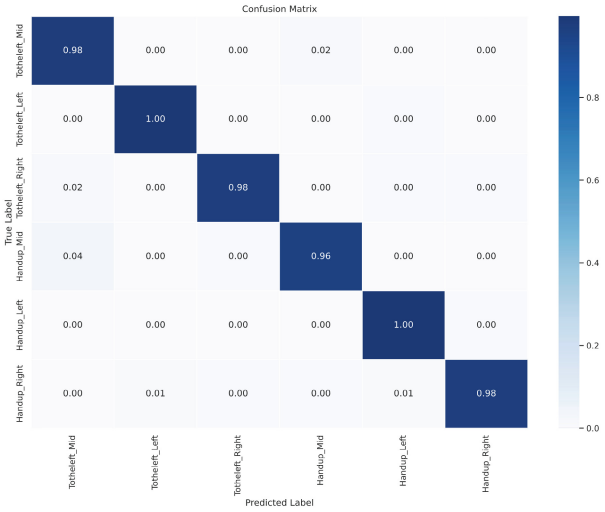


Fig. 4. Confusion matrix of the BiGRU.

The confusion matrix depicted in Fig. 4 provides additional insight into the BiGRU model's classification behavior. Nearly all diagonal entries approach perfect values, while off-diagonal elements remain minimal. The model demonstrates a strong ability to distinguish between all activity types and location categories, with very few misclassifications. Such reliability is especially valuable in real-world scenarios. Accurate joint inference of both activity and location is essential for context-sensitive systems and intelligent environment applications.

These findings support the hypothesis that BiGRU networks are particularly effective for analyzing WiFi CSI signals. This effectiveness arises from several key properties: (1) they efficiently model both brief and extended temporal dependencies introduced by human movement patterns in the signal; (2) bidirectional processing allows them to interpret full activity sequences by considering context in both directions; (3) they maintain lower computational demands due to fewer parameters compared to BiLSTM architectures; and (4) they deliver robust, consistent outcomes across varying data distributions. The slight yet steady improvement of BiGRU over the unidirectional GRU (98.33% versus 98.13%) confirms that the added computational overhead of bidirectional processing is warranted. It enables a richer interpretation of temporal structures within CSI data.

5 Conclusion and Future Works

This study has introduced a unified framework for jointly recognizing human activities and estimating indoor positions using WiFi CSI data. The system is built upon the BiGRU neural network architecture. Our work tackles a key problem in WiFi-based human sensing. Specifically, we show that both activity types

and spatial locations can be inferred simultaneously from identical CSI measurements. This removes the requirement for independent detection pipelines.

The proposed system is trained and evaluated on the WiARL-UIT dataset. This dataset is the first publicly accessible WiFi CSI collection that incorporates wireless timing synchronization designed for joint recognition tasks. It comprises 1,500 CSI samples spanning six unique activity-location pairings, offering a well-balanced basis for deep learning model development and assessment. A systematic preprocessing pipeline turned raw OFDM signals into clean, normalized time-series inputs. To get the data ready for use by a neural network, this pipeline included CSI extraction, amplitude correction, and value normalization.

The experimental results show that the BiGRU design works well. It had an F1-score of 98.33% (±0.30%), the smallest loss of 0.09 (±0.04), and an accuracy of 98.33% (±0.30%). The BiGRU outperformed several reference designs, such as CNN1D (90.80%), LSTM (90.00%), BiLSTM (95.67%), and the single-direction GRU (98.13%). These results support our hypothesis: integrating bidirectional sequence analysis with efficient gating mechanisms produces optimal feature learning from WiFi CSI data. The very minimal standard deviation across cross-validation repetitions shows that the behavior is steady and reliable. For real-world system deployment, this kind of consistency is very important.

The confusion matrix analysis shows that the model is quite good at classifying all activity-location groups. The diagonal values are almost perfect, and the misclassification rates are still quite low. These findings show that the BiGRU design can accurately capture temporal correlations in CSI signal variations caused by human mobility, both over short and long time periods. In the meantime, bidirectional processing enables the model to understand entire activity sequences by using input from both before and after the activity.

The results of this study are promising, but many areas need further research. First, the dataset needs to be expanded to include more activities, other locations, and a greater variety of people. This expansion would allow for a more comprehensive assessment of the framework's generalizability across diverse individuals and contexts. Second, the current method mostly uses amplitude data. Future work should explore more advanced methods for incorporating phase measurements via multimodal integration to improve recognition performance. Third, problems that make it hard to put these ideas into practice need to be solved. These include changes in the environment, interference from many users, and changes in the properties of wireless channels.

Also, we should explore ways to reduce model size, such as network pruning, weight quantization, and knowledge distillation. These techniques enable real-time inference on devices with minimal computing power and on embedded platforms, making them suitable for use in IoT environments. Finally, sensor fusion of WiFi CSI data with other sensing technologies, like accelerometers or radar systems, could lead to more reliable and accurate recognition frameworks. This is especially important in tough conditions because depending on just one sense modality might not be enough.

Acknowledgments. This research budget was allocated by University of Phayao; Thailand Science Research and Innovation Fund (Fundamental Fund 2026) with Grant No. 2255/2568; National Science, Research and Innovation Fund (NSRF); and King Mongkut's University of Technology North Bangkok with (Project no. KMUTNB-FF-69-B-02).

References

1. Al Mudawi, N., et al.: IoT powered RNN for improved human activity recognition with enhanced localization and classification. Sci. Rep. **15**(1), 10328 (2025). https://doi.org/10.1038/s41598-025-94689-5
2. Al-qaness, M.A.A., et al.: Channel state information from pure communication to sense and track human motion: a survey. Sensors **19**(15) (2019). https://doi.org/10.3390/s19153329
3. Bibbò, L., Carotenuto, R., Della Corte, F.: An overview of indoor localization system for human activity recognition (har) in healthcare. Sensors **22**(21) (2022). https://doi.org/10.3390/s22218119
4. Chary, V.S.K., Venkateswarlu, B.S.S., Vemuri, S., Pasupuleti, V.N.S.S., Burra, V.B., Tumuluru, P.: Activity identification via Wi-Fi channel state information with neural networks (2024). https://doi.org/10.21203/rs.3.rs-4794270/v1, preprint on Research Square
5. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder–decoder approaches. In: Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M. (eds.) Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/W14-4012, https://aclanthology.org/W14-4012/
6. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (GRU) neural networks. In: IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600 (2017). https://doi.org/10.1109/MWSCAS.2017.8053243
7. Elkelany, A., Ross, R., Mckeever, S.: WiFi-based human activity recognition using attention-based BiLSTM. In: Longo, L., O'Reilly, R. (eds.) AICS 2022. CCIS, vol. 1662, pp. 121–133. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26438-2_10
8. He, Z., Deng, K., Gong, J., Zhou, Y., Wang, D.: Transfer learning-enhanced instantaneous multi-person indoor localization by CSI (2024). https://arxiv.org/abs/2403.01153
9. Huang, S., et al.: WiMANS: a benchmark dataset for WiFi-based multi-user activity sensing. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024, Part XLII. LNCS, vol. 15100, pp. 72–91. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72946-1_5
10. Lynn, H.M., Pan, S.B., Kim, P.: A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks. IEEE Access **7**, 145395–145405 (2019). https://doi.org/10.1109/ACCESS.2019.2939947
11. Seow, M.J., Asari, V.: Recurrent neural network as a linear attractor for pattern association. IEEE Trans. Neural Netw. **17**(1), 246–250 (2006). https://doi.org/10.1109/TNN.2005.860869

12. Sohaib, S., Bokhari, S.M., Shafi, M., Alhashmi, A.: A novel approach for joint indoor localization and activity recognition using a hybrid CNN-GRU and MRF framework. PLOS ONE **20**(8), 1–18 (2025). https://doi.org/10.1371/journal.pone.0328181
13. Thanh, P.V.L., Ph., Q.N., Nguyen, D., Khuc, B., Gelgor, A., T.H., P.N.: Public dataset for simultaneous human activity recognition and localization using WiFi signals. In: 2024 International Conference on Electrical Engineering and Photonics (EExPolytech), pp. 206–209 (2024). https://doi.org/10.1109/EExPolytech62224.2024.10755579
14. Uçkun, F.A., Özer, H., Nurbaş, E., Onat, E.: Direction finding using convolutional neural networks and convolutional recurrent neural networks. In: 2020 28th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2020). https://doi.org/10.1109/SIU49456.2020.9302448
15. Wakili, A.A., Asaju, B.J., Jung, W.: Evaluating BiLSTM and CNN+GRU approaches for human activity recognition using WiFi CSI data (2025). https://arxiv.org/abs/2506.11165
16. Yousaf, J., et al.: Through-the-wall human activity recognition using radar technologies: a review. IEEE Open J. Anten. Propag. **5**(6), 1815–1837 (2024). https://doi.org/10.1109/OJAP.2024.3459045
17. Zhang, Y., Wang, X., Wen, J., Zhu, X.: WiFi-based non-contact human presence detection technology. Sci. Rep. **14**(1), 3605 (2024). https://doi.org/10.1038/s41598-024-54077-x

Efficient Satisfiability-Based Approach for Mining Partially-Ordered Sequential Rules

Abdelhamid Boudane(✉), Lamia Sadeg, and Mohamed Maiza

LMTO/UER-ISI, Ecole Militaire Polytechnique, Algiers, Algeria
{abdelhamid.boudane,lamia.sadeg,mohamed.maiza}@emp.mdn.dz

Abstract. Sequential rule mining is a key task in data mining that aims to discover temporal relationships between itemsets within sequence data. Among the different types of sequential rules, partially ordered sequential rules (*POSRules*) have gained attention due to their ability to capture precedence relationships without enforcing strict ordering within itemsets. However, existing approaches for mining *POSRules* often suffer from combinatorial explosion, high memory consumption, and limited flexibility in handling additional constraints.

In this paper, we propose a novel satisfiability (SAT)-based approach for extracting *POSRules*. The proposed method encodes the mining problem as a propositional formula whose models correspond to valid rules. To improve scalability, we introduce a decomposition strategy that partitions the original problem into smaller and more manageable subproblems, whose combined solutions yield the complete set of rules.

An extensive experimental evaluation conducted on several real-world datasets demonstrates that the proposed approach outperforms state-of-the-art algorithms in terms of efficiency and scalability. These results highlight the potential of declarative methods as a powerful alternative for sequential rule mining.

Keywords: Data mining · Sequential rules · Satisfiability · Propositional logic

1 Introduction

Data mining is a fundamental field that focuses on extracting hidden patterns and relationships from data. These relationships provide useful insights that help better understand the behavior of the studied phenomena. The type of the extracted knowledge depends on both the nature of the data and the objectives of the analysis. Among the various data mining tasks, sequential pattern mining plays an important role when dealing with ordered data. It has shown strong effectiveness in many application domains. For example, in e-commerce, it helps model customer purchasing behavior and improve recommendation systems by analyzing frequent purchase and click sequences [5,8]. In healthcare, it is used to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 129–145, 2027.
https://doi.org/10.1007/978-3-032-30494-0_9

discover temporal relationships in clinical data, supporting outcome prediction and decision-making [17]. These examples illustrate the ability of sequential pattern mining to capture meaningful temporal dependencies in complex sequence data.

Within different types of patterns in sequence data, sequential rules are of particular interest. They are expressed as implications of the form $X \rightarrow Y$ and can be used to predict future events in sequences. The literature distinguishes two main categories of sequential rules. The first category includes rules where both the antecedent X and the consequent Y are sequential patterns [20]. The second category consists of rules defined between two unordered sets of items [9,10,12].

In this paper, we focus on the second category, namely partially ordered sequential rules ($POSRules$) [10], which extend association rule mining by discovering relationships of the form $X \rightarrow Y$, where the items in X precede those in Y within sequences. In this setting, no ordering is imposed among the items within X or within Y; however, all items in X must occur before those in Y, hence the term partially ordered. The relevance of such rules is typically evaluated using support and confidence measures, which respectively quantify the frequency of a rule in the dataset and the strength of the implication between its antecedent and consequent.

Several algorithms have been developed to efficiently mine $POSRules$. Early methods, such as CMDeo [9], use an Apriori-based level-wise strategy that generates and tests candidate rules. While effective on small datasets, this approach does not scale well due to the large number of candidates. To overcome this limitation, CMRules [10] first extracts frequent itemsets and generates association rules, which are then filtered to meet sequential constraints. This reduces the search space compared to CMDeo, but its performance still depends on the number of generated rules. More recent approaches use pattern-growth techniques. For example, RuleGrowth [12] avoids candidate generation by directly extending rules on both sides, improving efficiency and scalability. Similarly, ERMiner [11] further improves performance using equivalence classes and advanced pruning, achieving faster runtime at the cost of higher memory use.

Despite these advances, existing approaches for mining $POSRules$ still face major challenges. They often suffer from combinatorial explosion, high memory requirements, and limited flexibility in handling additional constraints. These limitations motivate the exploration of alternative paradigms that can improve both efficiency and expressiveness in sequential rule mining.

Declarative approaches, such as satisfiability (SAT) and constraint programming, have been successfully applied to many data mining tasks [7], including association rule mining [3,4,14], community detection [16], and frequent sequence mining [1]. However, to the best of our knowledge, they have not yet been used for extracting partially ordered sequential rules. In this paper, we propose a SAT-based approach for mining $POSRules$. The method encodes the problem as a propositional formula, where each model corresponds to a valid rule. It also introduces a decomposition strategy that breaks the problem into

smaller subproblems, whose combined solutions form the complete set of rules. An extensive experimental evaluation on real-world datasets demonstrates that the proposed approach outperforms state-of-the-art methods in both efficiency and scalability.

2 Preliminaries and Problem Statements

2.1 Partially Ordered Sequential Rules Mining Problem

Let Σ be a finite non empty set of symbols, called *items*. An *itemset* X over Σ is defined as a subset of Σ, i.e., $X \subseteq \Sigma$. From now, we use the capital letters X, Y, K, etc. to represent itemsets and lowercase letters a, b, c, etc. to denote the elements of Σ.

A *sequence* S_i is a pair (i, S) where i is the sequence identifier and S is an ordered list of itemsets , i.e., $S = \{X_1, ..., X_n\}$. A *sequence database* $\mathcal{SD}$ is a finite non empty set of sequences where each sequence identifier refers to a unique list of itemsets.

An itemset X is covered by a sequence S if each item appearing in X appears in S. Then, the *cover* of an itemset X in a sequence database $\mathcal{SD}$, denoted $Cov(X, \mathcal{SD})$, is defined as $\{i \in \mathbb{N} \mid (i, S) \in \mathcal{SD}\ and\ X \subseteq \bigcup_{I \in S} I\}$.

The *support* of X in $\mathcal{SD}$, denoted $Supp(X, \mathcal{SD})$, corresponds to the cardinality of $Cov(X, \mathcal{SD})$, i.e., $Supp(X, \mathcal{SD}) =\mid Cov(X, \mathcal{SD}) \mid$.

A *partially-ordered sequential rule* ($POSRule$), is a pattern of the form $X \rightarrow Y$ where X and Y are two disjoint itemsets called respectively the antecedent and the consequence of the rule. the relevance of such rules is defined using two well known predicates namely support and confidence. We say that a POSRule $X \rightarrow Y$ is coverd by a sequence $S = \{I_1, ..., I_n\}$ and we note $(X \rightarrow Y) \sqsubseteq S$ iff $\exists\ i,\ j,\ k,\ l \in \mathbb{N} \mid\ 1 \leq i \leq j < k \leq l \leq n$ and $X \subseteq \{\bigcup_{m=i}^{m=j} I_m\}$ and $Y \subseteq \{\bigcup_{m=k}^{m=l} I_m\}$. The constraint $j < k$ represents the fact that the items of Y must appear after the items of X while there is no order on the elements of X and Y.

The *cover* of a $POSRule$ $X \rightarrow Y$ in a sequential database $\mathcal{SD}$, is defined as $\mathcal{C}ov(X \rightarrow Y, \mathcal{SD}) = \{i \in \mathbb{N} \mid (i, S) \in \mathcal{SD}\ and\ (X \rightarrow Y) \sqsubseteq S_i\}$. Then, the *support* of a $POSRule$ $X \rightarrow Y$ in $\mathcal{SD}$, is defined as $Supp(X \rightarrow Y, \mathcal{SD}) = \frac{|\mathcal{C}ov(X \rightarrow Y, \mathcal{SD})|}{|\mathcal{SD}|}$. Such support measure determines the frequency of a candidate $POSRule$ in a given sequential database $\mathcal{SD}$.

The *confidence* of $X \rightarrow Y$ in $\mathcal{S}$, defined as $\mathcal{C}onf(X \rightarrow Y, \mathcal{S}) = \frac{Supp(X \rightarrow Y, \mathcal{SD})}{Supp(X, \mathcal{SD})}$, represents the conditional probability of the appearance of Y after a given X in $\mathcal{SD}$. When there is no ambiguity, we omit to mention the sequential database $\mathcal{SD}$, and we simply note $Supp(X \rightarrow Y)$ and $\mathcal{C}onf(X \rightarrow Y)$. Note that support (respectively confidence) measures can be expressed as a percentage by multiplying it by 100.

This work addresses the problem of mining partially ordered sequential rules, formalized as follows:

Definition 1 (Problem of mining POSRules). *Given a sequence database $\mathcal{SD}$, a minimum support threshold λ and a minimum confidence threshold θ, the problem of mining POSRules consists in finding all the pairs of itemsets (X,Y) so that the following constraints are satisfied:*

- $X \cap Y = \emptyset$.
- $Supp(X \rightarrow Y) \geq \lambda$.
- $\mathcal{C}onf(X \rightarrow Y) \geq \theta$

Formally, the set of partially ordered rules is defined as follows:

$$POSRules(\mathcal{SD}, \lambda, \theta) = \{X \rightarrow Y \mid X, Y \subseteq \Sigma,\ Supp(X \rightarrow Y) \geq \lambda, \\ and\ \mathcal{C}onf(X \rightarrow Y) \geq \theta\}$$

Example 1. Consider the sequence database $\mathcal{SD}$, represented in Table 1, containing four sequences over the set of items $\Sigma = \{a, b, c, d, e, f, g\}$. For a minimum support threshold $\lambda = 0.5$ and a minimum confidence threshold $\theta = 0.7$ there are five partially ordered rules illustrated in Table 2. For example, for the rule $\{b\} \rightarrow \{f\}$ we have $\mathcal{C}ov(\{b\} \rightarrow \{f\}, \mathcal{SD}) = \{1, 2, 4\}$ and $\mathcal{C}ov(\{b\}, \mathcal{SD}) = \{1, 2, 3, 4\}$. Therefore, $Supp(\{b\} \rightarrow \{f\}, \mathcal{S}) = 3/4$ and $\mathcal{C}onf(\{b\} \rightarrow \{f\}, \mathcal{S}) = 3/4$.

Table 1. A Database $\mathcal{SD}$

S_{id}	*Sequences*
1	ab, g, f, c, e
2	de, bc, a, f, b
3	a, g, b, ac
4	b, fe

Table 2. POSRules

Rules	*Supp*	$\mathcal{C}onf$
$\{b\} \rightarrow \{f\}$	3/4	3/4
$\{g\} \rightarrow \{c\}$	2/4	2/2
$\{a, g\} \rightarrow \{c\}$	2/4	2/2
$\{a, b, g\} \rightarrow \{c\}$	2/4	2/2
$\{b, g\} \rightarrow \{c\}$	2/4	2/2

2.2 Satisfiability Problem in Propositional Logic

We introduce in the following the fundamental concepts of propositional logic. Indeed, We define at first the syntax and the semantics of propositional logic and we present at second the satisfiability problem in such logic.
A propositional variable x is a Boolean input variable of a truth function that takes values in {0,1}. 0 and 1 are used to denote $False$ ($\bot$) and $True$ ($\top$) values. Let $\mathcal{B}$ be the set of Boolean variables. The set of propositional formulas denoted by $\mathcal{F}$ is defined from $\mathcal{B}$, $\bot$, $\top$ and the logical connectives ($\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$) so that:

- A propositional variable $x \in \mathcal{B}$, $\bot$, and $\top$ are formulas;
- If ϕ and ψ are propositional formulas, then $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$ and $\phi \leftrightarrow \psi$ are also propositional formulas.

We use $\mathcal{V}(\phi)$ to denote the set of Boolean variables appearing in formula ϕ.
Basically, logical reasoning solvers take input formulas in normal forms like

conjunctive normal form (CNF). A CNF is a conjunction of clauses. A clause is a disjunction of literals where a literal is a positive propositional variable x or its negation $\neg x$.
An interpretation $\mathcal{I}$ of a propositional formula ϕ is a function from $\mathcal{V}(\phi)$ to $\{0,1\}$ which associates to each Boolean variable $x \in \mathcal{V}(\phi)$ a value $\mathcal{I}(x) \in \{0,1\}$. An interpretation $\mathcal{I}$ that satisfies a formula ϕ ($\mathcal{I}(\phi) = 1$) is called model of ϕ.
Given a CNF formula ϕ, The Satisfiability (SAT) problem consists on deciding if ϕ admits a model or not. SAT is a fundamental problem in computational complexity theory because it was the first problem to be proved NP-Complete [6]. Its worth noting that the mostly used satisfiability solvers takes as input a CNF formula and any propositional formula could be transformed to an equisatisfiable CNF formula in polynomial time and space [19].

3 SAT Encoding of Mining Partially Ordered Sequential Rules

In this section, we propose a SAT-based encoding for the problem of mining partially ordered sequential rules (*POSRules*). The approach consists in representing the mining task as a propositional formula in conjunctive normal form (CNF), such that each satisfying assignment corresponds to a valid *POSRule*. This encoding relies on the definition of propositional variables along with a set of constraints that precisely capture the semantics of the problem. We begin by introducing the variables used to represent the different components and positions of a sequential rule. We then present the constraints required to ensure a correct and complete modeling of the *POSRules* mining problem.

3.1 Variables Definition

Given a set Σ of n items, a sequential database $\mathcal{SD} = \{(1, S_1), ..., (m, S_m)\}$ where $\forall i \in \{1, ..., m\}, S_i = \{I_1 \subseteq \Sigma, ..., I_k \subseteq \Sigma\}$, a minimum support threshold λ and a minimum confidence threshold θ. We define two sets of Boolean variables $\{x_a \mid a \in \Sigma\}$ and $\{y_a \mid a \in \Sigma\}$ to represent the items of a candidate sequential rule $X \rightarrow Y$. In fact, true variables from the two sets, in a given model $\mathcal{M}$ of the SAT encoding, represent the *POSRule* $X = \{x_{a \in \Sigma},\ x_a = 1\} \rightarrow Y = \{y_{a \in \Sigma},\ y_a = 1\}$.

We associate also to each sequence identifier i two propositional variables p_i and q_i respectively. This variables allows us to model the fact that a sequence (i, S) in $\mathcal{SD}$ covers the itemset X and the candidate rule $X \rightarrow Y$ respectively. Formally, in a model $\mathcal{M}$ of the considered encoding, the two sets $\{i \in \{1, ..., m\} \mid \mathcal{M}(p_i) = 1\}$ and$\{\{i \in \{1, ..., m\} \mid \mathcal{M}(q_i) = 1\}\}$ represent respectively $Cov(X, \mathcal{SD})$ and $Cov(X \rightarrow Y, \mathcal{SD})$.

For example, the corresponding Boolean variables of the dataset $\mathcal{SD}$ presented in Table 1 are $X = \{x_a, x_b, x_c, x_d, x_e, x_f, x_g\}$, $Y = \{y_a, y_b, y_c, y_d, y_e, y_f, y_g\}$, $Cov(X, \mathcal{SD}) = \{p_1, p_2, p_3, p_4\}$, and $Cov(X \rightarrow Y, \mathcal{SD}) = \{q_1, q_2, q_3, q_4\}$.

3.2 Constraints Formulation

We now introduce the set of constraints used to model the *POSRules* extraction problem. Recall that the validation of a candidate rule $X \rightarrow Y$ relies on the computation of support and confidence measures. In the proposed encoding, these measures require the definition of constraints that capture both the cover of the antecedent, denoted $\mathcal{C}ov(X, \mathcal{SD})$, and the cover of the rule, denoted $\mathcal{C}ov(X \rightarrow Y, \mathcal{SD})$. To this end, we encode the occurrence of the candidate rule $X \rightarrow Y$ and its antecedent X in each sequence S_i using Boolean variables q_i and p_i, respectively. These variables are then used within linear constraints to express the support and confidence conditions.

We now present the different constraints that model the problem of mining *POSRules*.

1. *The two sides of the rule are not empty (i.e. $X \neq \emptyset$ and $Y \neq \emptyset$):* this constraint is expressed by two clauses (see formula (1)) that enforce the candidate rule to have at least one item in X and at least one item in Y. It avoids candidate rules that doesn't have conditions or consequences like ($\rightarrow y_b$ *or* $x_a \rightarrow$).

$$(\bigvee_{a \in \Sigma} x_a) \wedge (\bigvee_{a \in \Sigma} y_a) \tag{1}$$

2. *An item can appear in at most one side of a candidate rule (i.e. $X \cap Y = \emptyset$):* This constraint (see Formula (2)) ensures that no item is shared between the antecedent X and the consequent Y, thereby eliminating invalid rules such as $ac \rightarrow ab$. More precisely, for each item $a \in \Sigma$, a clause is introduced to enforce that a cannot simultaneously belong to both X and Y.

$$\bigwedge_{a \in \Sigma} (\neg x_a \vee \neg y_a) \tag{2}$$

3. *A sequence S_i covers the left side of a candidate rule $X \rightarrow Y$:* This constraint captures whether the itemset X appears in the sequence S_i. Specifically, Formula (3) enforces the variable p_i to be false if and only if at least one item of X does not appear in S_i.

$$\neg p_i \leftrightarrow \bigvee_{a \notin s_i} x_a \tag{3}$$

4. *The itemset Y appears after the itemset X in a sequence S_i:* To model this constraint, an additional variable z_i is introduced to ensure that all items in X occur before those in Y within S_i. This condition is captured by Formula (4), which enforces that each item $y \in Y$ appears after the first occurrence of every item $x \in X$.
   In other words, if there exists an item $a \in X$ whose first occurrence is after the last occurrence of an item $b \in Y$, then the candidate rule $X \rightarrow Y$ violates the ordering constraint in sequence S_i, and the variable z_i is set to false.

$$\forall S_i \in \mathcal{SD}, \bigwedge_{a \in s_i,\ b \in s_i} z_i \rightarrow (x_a \rightarrow \neg y_b); \quad FOcc(a, s_i) \geq LOcc(b, s_i). \tag{4}$$

Note that $FOcc(a, s_i)$ (respectively $LOcc(b, s_i)$) is a function that returns the position of the first (respectively the last) occurrence of the item a (respectively b) in a sequence S_i. Moreover, formula (4) can be simply rewritten as a conjunction of clauses using the equivalence $A \rightarrow B \equiv \neg A \vee B$ as follows:

$$\forall S_i \in \mathcal{SD}, \bigwedge_{a \in s_i,\ b \in s_i} (\neg z_i \vee \neg x_a \vee \neg y_b); \quad FOcc(a, s_i) \geq LOcc(b, s_i).$$

5. *A sequence S_i covers the candidate rule $X \rightarrow Y$:* This constraint uses the variable q_i to represent whether a candidate rule $X \rightarrow Y$ is covered by a sequence S_i. For this to hold, both X and Y must appear in S_i, and the items of X must precede those of Y. If any of these conditions is not satisfied, the variable q_i is set to false. Formula (5) encodes this constraint by combining the variable p_i, which represents the cover of X, and the variable z_i, which enforces the ordering constraint. In addition, the subformula $\bigvee_{a \notin S_i} y_a$ is used to capture the cover of Y.

$$\neg q_i \leftrightarrow \neg p_i \vee \bigvee_{a \notin s_i} y_a \vee \neg z_i \tag{5}$$

6. *Minimum support:* This constraint encodes the fact that a candidate rule $X \rightarrow Y$ should have a minimum support λ in the database $\mathcal{SD}$ to be a valid *POSRule*. In fact, it should be covered by at least λ sequences. Such cover tests are captured by variables q_i from formula (5). Then, the sum of true q_i variables gives us the support value of the candidate rule $X \rightarrow Y$ as illustrated in formula (6).

$$\sum_{i=1}^{i=m} q_i \geq \lambda \tag{6}$$

7. *Minimum confidence:* To express this constraint, it is necessary to compute both the support of the itemset X and the support of the candidate rule $X \rightarrow Y$. These values are obtained by summing the variables p_i and q_i, which are defined by the cover constraints in Formulas (3) and (5), respectively. The minimum confidence constraint is then modeled as a 0/1 linear inequality, as shown in Formula (7).

$$\frac{\sum_{i=1}^{i=m} q_i}{\sum_{i=1}^{i=m} p_i} \geq \theta \tag{7}$$

By considering that the confidence threshold θ takes values between 0 and 1, we can rewrite the constraint 7 as follows:

$$100 * \sum_{i=1}^{i=m} q_i - 100 * \theta * \sum_{i=1}^{i=m} p_i \geq 0 \tag{8}$$

It is worth noting that the two formulas (6) and (8) are called cardinality and pseudo Boolean constraints. They could be translated efficiently into CNF formulas using one of the literature encodings ([2,18]).

We use $\Phi_{POSRules}(\mathcal{SD}, \lambda, \theta)$ to denote the encoding corresponding to the conjunction of the propositional formulas (1), (2), (3), (4), (5), (6), and (8).

Proposition 1. *Let $\mathcal{SD}$ a sequential database, λ a minimum support threshold, θ a minimum confidence threshold and $\mathcal{I}$ an interpretation of $\Phi_{POSRules}(\mathcal{SD}, \lambda, \theta)$. Then, $\mathcal{I}$ is a model of $\Phi_{POSRules}(\mathcal{SD}, \lambda, \theta)$ iff $X = \{a \in \Sigma | \mathcal{I}(x_a) = 1\} \rightarrow Y = \{b \in \Sigma | \mathcal{I}(y_b) = 1\}$ is a POSRule with $\mathcal{C}ov(X \rightarrow Y, \mathcal{SD}) = \{i \in \{1..m\} | \mathcal{I}(q_i) = 1\}$.*

Example 2. Let us consider the sequencial database of Table 1. The following formula encodes the problem of mining Partially Ordered Rules with a minimum support threshold $\lambda = 0.5$ (at least 2 sequences covers the rule) and a minimum confidence threshold $\theta = 0.5$:

$$
\begin{aligned}
&(x_a \vee x_b \vee x_c \vee x_d \vee x_e \vee x_f \vee x_g) \wedge (y_a \vee y_b \vee y_c \vee y_d \vee y_e \vee y_f \vee y_g) \wedge \\
&(\neg x_a \vee \neg y_a) \wedge (\neg x_b \vee \neg y_b) \wedge (\neg x_c \vee \neg y_c) \wedge (\neg x_d \vee \neg y_d) \wedge (\neg x_e \vee \neg y_e) \wedge \\
&(\neg x_f \vee \neg y_f) \wedge (\neg x_g \vee \neg y_g) \wedge \\
&(\neg p_1 \leftrightarrow x_d) \wedge \\
&(\neg p_2 \leftrightarrow x_g) \wedge \\
&(\neg p_3 \leftrightarrow x_d \vee x_e \vee x_f) \wedge \\
&(\neg p_4 \leftrightarrow x_a \vee x_c \vee x_d \vee x_g) \wedge \\
&(z_1 \rightarrow (x_a \rightarrow \neg y_b)) \wedge (z_1 \rightarrow (x_b \rightarrow \neg y_a)) \wedge (z_1 \rightarrow (x_c \rightarrow \neg y_a)) \wedge \\
&(z_1 \rightarrow (x_c \rightarrow \neg y_b)) \wedge (z_1 \rightarrow (x_c \rightarrow \neg y_f)) \wedge (z_1 \rightarrow (x_c \rightarrow \neg y_g)) \wedge \\
&(z_1 \rightarrow (x_e \rightarrow \neg y_a)) \wedge (z_1 \rightarrow (x_e \rightarrow \neg y_b)) \wedge (z_1 \rightarrow (x_e \rightarrow \neg y_c)) \wedge \\
&(z_1 \rightarrow (x_e \rightarrow \neg y_f)) \wedge (z_1 \rightarrow (x_e \rightarrow \neg y_g)) \wedge (z_1 \rightarrow (x_f \rightarrow \neg y_a)) \wedge \\
&(z_1 \rightarrow (x_f \rightarrow \neg y_b)) \wedge (z_1 \rightarrow (x_f \rightarrow \neg y_g)) \wedge \\
&(z_2 \rightarrow (x_a \rightarrow \neg y_c)) \wedge (z_2 \rightarrow (x_a \rightarrow \neg y_d)) \wedge (z_2 \rightarrow (x_a \rightarrow \neg y_e)) \wedge \\
&(z_2 \rightarrow (x_b \rightarrow \neg y_c)) \wedge (z_2 \rightarrow (x_b \rightarrow \neg y_d)) \wedge (z_2 \rightarrow (x_b \rightarrow \neg y_e)) \wedge \\
&(z_2 \rightarrow (x_c \rightarrow \neg y_d)) \wedge (z_2 \rightarrow (x_c \rightarrow \neg y_e)) \wedge (z_2 \rightarrow (x_d \rightarrow \neg y_e)) \wedge \\
&(z_2 \rightarrow (x_e \rightarrow \neg y_d)) \wedge (z_2 \rightarrow (x_f \rightarrow \neg y_a)) \wedge (z_2 \rightarrow (x_f \rightarrow \neg y_c)) \wedge \\
&(z_2 \rightarrow (x_f \rightarrow \neg y_d)) \wedge (z_2 \rightarrow (x_f \rightarrow \neg y_e)) \wedge \\
&(z_3 \rightarrow (x_b \rightarrow \neg y_a)) \wedge (z_3 \rightarrow (x_b \rightarrow \neg y_g)) \wedge (z_3 \rightarrow (x_c \rightarrow \neg y_a)) \wedge \\
&(z_3 \rightarrow (x_c \rightarrow \neg y_b)) \wedge (z_3 \rightarrow (x_c \rightarrow \neg y_g)) \wedge \\
&(z_4 \rightarrow (x_e \rightarrow \neg y_b)) \wedge (z_4 \rightarrow (x_e \rightarrow \neg y_f)) \wedge (z_4 \rightarrow (x_f \rightarrow \neg y_b)) \wedge \\
&(z_4 \rightarrow (x_f \rightarrow \neg y_e)) \wedge \\
&(\neg q_1 \leftrightarrow \neg p_1 \vee y_d \vee \neg z_1) \wedge \\
&(\neg q_2 \leftrightarrow \neg p_2 \vee y_g \vee \neg z_2) \wedge \\
&(\neg q_3 \leftrightarrow \neg p_3 \vee y_d \vee y_e \vee y_f \vee \neg z_3) \wedge \\
&(\neg q_4 \leftrightarrow \neg p_4 \vee y_a \vee y_c \vee y_d \vee y_g \vee \neg z_4) \wedge \\
&(q_1 + q_2 + q_3 + q_4 \geq 2) \wedge \\
&(10 \times (q_1 + q_2 + q_3 + q_4) - 5 \times (p_1 + p_2 + p_3 + p_4) \geq 0)
\end{aligned}
$$

Its worth noticing that useful user constraints such as minimum (respectively maximum) conditions α and minimum (respectively maximum) consequences β in a $POSRule\ X \rightarrow Y$ could easily be modeled using the following formulas:

– *Maximum (respectively Minimum) Antecedent:* Formula (9) expresses that the number of items in the antecedent X of a *POSRule* shouldn't exceed (respectively has less than) α.

$$\sum_{a \in \Sigma} x_a \leq (resp. \geq) \quad \alpha \tag{9}$$

– *Maximum (respectively Minimum) Consequence:* Formula (10) expresses that the number of items in the consequence Y of a *POSRule* $X \rightarrow Y$ shouldn't exceed (respectively has less than) β.

$$\sum_{a \in \Sigma} y_a \leq (resp. \geq) \quad \beta \tag{10}$$

4 Two-Step Decomposition Approach for SAT Based Enumeration of POSRules

Extracting *POSRules* using the SAT-based approach described in Sect. 3 involves enumerating all models of the propositional formula $\Phi_{POSRules}(\mathcal{SD}, \lambda, \theta)$. The efficiency of this approach relies on how quickly these models can be enumerated. However, for large sequential databases $\mathcal{SD}$, the encoding can produce formulas with hundreds of thousands of variables and millions of clauses, which can significantly reduce both efficiency and scalability. In addition, the number of *POSRules* is typically very large, further slowing down the model enumeration process.

To address these challenges, we propose a decomposition-based approach to improve both scalability and efficiency of the proposed approach. This method splits the original dataset into smaller partitions $\mathcal{SD}_i$ and encodes each partition separately as $\Phi_{POSRules}(\mathcal{SD}_i, \lambda, \theta)$. Each partition produces a smaller formula with fewer items, which reduces computational complexity and allows faster model enumeration, improving the performance of the SAT-based approach introduced in Sect. 3.

We now describe the two-step decomposition approach that we propose for enumerating *POSRules* using SAT (see Algorithm 1). First, all items in the input sequential dataset $\mathcal{SD}$ are sorted in ascending order of support and stored in a vector $\mathcal{V}$. In the first decomposition step, we iterate over each item a in $\mathcal{V}$ and project $\mathcal{SD}$ into an intermediate partition $\mathcal{SD}'$ containing only sequences where a appears (iterations 25).

In the second step, for each item $a' \neq a$ appearing in $\mathcal{SD}'$, we project $\mathcal{SD}'$ into a final partition $\mathcal{SD}''$ containing sequences where both a and a' occur (iterations 78). Each final partition $\mathcal{SD}''$ is then encoded as $\Phi_{POSRules}(\mathcal{SD}'', \lambda, \theta)$, and all models of the resulting CNF formula are enumerated, with each model representing a *POSRule*. Variables x_a and $x_{a'}$ are set to true in the encoding of $\mathcal{SD}''$, while all items in $\mathcal{T}$ that have already been processed in previous iterations are set to false (iterations 812). This propagation prevents redundant

enumeration, ensuring that for each $\mathcal{SD}''$ we only enumerate rules of the form $X = aa'* \rightarrow Y$, where $*$ represents a subset of yet-unprocessed items appearing witnin $\mathcal{SD}''$.

At the end of the first step (iterations 1315), we handle the special case of *POSRules* of the form $a \rightarrow Y$ by encoding $\mathcal{SD}'$ with x_a set to true and all previously processed items set to false. All *POSRules* enumerated from different partitions are disjoint, and their union produces exactly the complete set of rules from the original dataset $\mathcal{SD}$. In the worst case, the number of generated encodings (i.e. partitions) is $O(|\Sigma|^2)$.

Finally, to improve efficiency, items with support less than λ are skipped, reducing the search space and accelerating the overall enumeration procedure. It is worth noting that the function *encode_cnf*() (lines 10 and 14) employs the encoding proposed in Sect. 3, whereas *enumerate_models*() (lines 11 and 15) relies on a slightly modified solver that adopts a DPLL-like procedure to enumerate all models of a given CNF formula [15].

Algorithm 1: Two-Step Decomposition-based *POSRules* Mining

Data: A sequence database $\mathcal{SD}$, Minimum support λ, Minimum confidence θ
Result: Set of all *POSRules* R

1 $R \leftarrow \emptyset$, $\Gamma \leftarrow \emptyset$, $V = \langle a_1 \dots a_n \rangle \leftarrow items(\mathcal{SD})$;
2 **for** $a \in V$ **do**
3 $\quad \mathcal{SD}' \leftarrow \{S_i \in \mathcal{SD} \mid a \in S_i\}$;
4 $\quad V' \leftarrow items(\mathcal{SD}') \setminus \{a\}$;
5 $\quad \Gamma \leftarrow \Gamma \cup \{a\}$;
6 $\quad \mathcal{T} \leftarrow \Gamma$
7 $\quad$ **for** $a' \in V'$ **do**
8 $\quad\quad \mathcal{SD}'' \leftarrow \{S_i \in \mathcal{SD}' \mid a' \in S_i\}$;
9 $\quad\quad \mathcal{T} \leftarrow \mathcal{T} \cup \{a'\}$;
10 $\quad\quad F = encode_cnf(\mathcal{T}, \mathcal{SD}'', \lambda, \theta)$
11 $\quad\quad R = R \cup enumerate_models(F)$;
12 $\quad\quad \mathcal{T} \leftarrow (\mathcal{T} \setminus \{a'\}) \cup \{\neg a'\}$;
13 $\quad$ **end**
14 $\quad F = encode_cnf(\mathcal{T}, \mathcal{SD}', \lambda, \theta)$
15 $\quad R = R \cup enumerate_models(F)$;
16 $\quad \Gamma \leftarrow (\Gamma \setminus \{a\}) \cup \{\neg a\}$;
17 **end**
18 **return** R;

5 Experimental Evaluation

To evaluate the proposed SAT-based approach for mining Partially Ordered Sequential Rules ($SPOSR$), we conducted a series of experiments on several real-world datasets. The performance of $SPOSR$ was compared with that of state-of-the-art specialized algorithms, namely CMRule [10], CMDEO [9],RuleGrowth

[12], and ERMiner [11], all implemented in the SPMF tool[1]. The evaluation focuses on two criteria: the execution time required to extract all the desired *POSRules*, and the peak memory consumption during the mining process. We implemented *SPOSR* on top of the well-known SAT solver ManySAT [13], which is written in C++. The enumeration of all models (i.e., rules) is performed using a DPLL-based procedure [15]. All experiments were conducted on a machine equipped with an Intel i5 processor, 16 GB of RAM, and running at 2.30 GHz under Ubuntu 18.04. The *SPOSR* solver, along with the datasets used in this study, is available at the following link[2] to facilitate reproducibility.

Datasets

In these experiments, we used real-world datasets obtained from the SPMF website[3]. To ensure a comprehensive evaluation, we selected a diverse range of datasets, from small ones such as *SIGN* to large-scale datasets like *kosarak*, which contains approximately 1 million sequences and about 41 thousand distinct items. Table 3 summarizes the main characteristics of each dataset, including the number of sequences (#Sequences), the number of items (#Items), and the average sequence length (AGV.Length).

Table 3. Datasets characteristics

Dataset	#Sequences	#Items	AGV.Length
SIGN	730	310	51.9
retail best	4383	10157	48.23
retail all	4383	41431	32.15
Leviathan	5834	9494	33.8
E-shop	24026	317	61.9
MSNBC	31790	18	13.2
Bible	36369	14576	21.6
BMS1	59601	60978	2.1
BMS2	77512	330286	4.62
Kosarak	990002	41271	8.1

Experimental setup

To evaluate the behavior of *SPOSR* for different minimum support thresholds λ, we set a time limit of 1000 s and a memory limit of 15 GB. We then measure

[1] https://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php.
[2] https://drive.google.com/drive/folders/1LtxhTnm93_skDtA0CbOzSyxHxYS1pyiO?usp=sharing.
[3] https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php.

the performance (specifically, enumeration time and memory usage) of the algorithms (CMRule, CMDeo, RuleGrowth, ERMiner, and $SPOSR$) as a function of λ, while keeping the minimum confidence fixed at $\theta = 80\%$.

Results

Table 4. Enumeration time evaluation

Dataset	Min-supp	#Rules	Execution Time(s)				
			CMRule	CMDEo	RuleGrowth	ERMiner	SPOSR
SIGN	15%	174509	TO	TO	123.38	59.96	**5.62**
	10%	1083281	TO	TO	672.21	358.12	**18.51**
	5%	20986080	TO	TO	TO	TO	**260.26**
Leviathan	1%	43	310.54	894.89	89.12	95.75	**7.73**
	0.5%	292	TO	TO	295.27	410.87	**17.13**
	0.1%	733487	TO	TO	TO	TO	**156.17**
MSNBC	0.8%	1616	87.89	198.03	70.73	62.32	**2.60**
	0.5%	3440	131.01	667.61	206.62	153.02	**2.89**
	0.2%	5240	205.23	TO	527.00	355.82	**3.24**
Bible	0.5%	142	80.19	TO	136.00	120.98	**27.13**
	0.1%	27173	TO	TO	TO	MO	**138.95**
	0.05%	2042009	TO	TO	TO	MO	**279.11**
BMS1	0.08%	2744	9.20	17.22	4.56	3.66	**2.92**
	0.06%	1123969	TO	TO	131.24	46.29	**7.93**
	0.057%	20446658	TO	TO	TO	MO	**37.78**
BMS2	0.02%	868973	TO	TO	373.38	309.74	**48.13**
	0.01%	6310015	TO	TO	TO	TO	**104.55**
	0.005%	62931480	TO	TO	TO	TO	**268.11**
Kosarak	0.7%	126	41.43	208.89	71.54	**31.94**	34.55
	0.5%	286	80.30	568.70	135.36	**55.65**	57.72
	0.3%	1362	227.06	TO	359.96	**138.48**	142.42
retail best	2%	28770	TO	TO	/	101.82	**8.90**
	1%	677297	TO	TO	/	MO	**58.76**
	0.6%	7100174	TO	TO	/	MO	**385.25**
retail all	0.9%	28770	TO	TO	/	**103.14**	106.48
	0.7%	297448	TO	TO	/	377.48	**208.75**
	0.5%	2713178	TO	TO	/	MO	**469.16**
E-shop	20%	1433	TO	TO	/	170.97	**17.21**
	15%	9666	TO	TO	/	MO	**39.29**
	10%	142424	TO	TO	/	MO	**269.64**

Table 4 reports the results of the experiments, showing the execution times of the state-of-the-art algorithms CMRule, CMDEO, ERMiner, RuleGrowth, as well as

the proposed approach *SPOSR* for enumerating *POSRules*. As expected, the execution time increases as the minimum support decreases, which also corresponds to an increase in the number of *POSRules*. Overall, *SPOSR* outperforms the other algorithms across all datasets, except for *kosarak*, where *ERMiner* achieves slightly better results. Additionally, *CMDEO* and *CMRule* exceed the time limit (TO) in most cases, particularly for datasets containing sequences of itemsets, such as *E-shop*, *retail best* or *retail all*. It is also important to note that *RuleGrowth* was excluded from experiments on datasets with sequences of itemsets, as it does not handle such datasets correctly.

Figure 1 presents a scatter plot comparing *SPOSR* with other algorithms in terms of execution time across all experiments. It clearly shows that *SPOSR* consistently outperforms the competing approaches on every dataset. For a more detailed analysis, Figs. 2 and 3 illustrate the behavior of all algorithms with respect to both execution time and memory consumption. These figures depict, for each dataset, the evolution of enumeration time, memory usage, and the number of extracted rules as a function of the minimum support threshold.

Overall, the results indicate that, for the vast majority of datasets, *SPOSR* significantly outperforms the other algorithms in both runtime and memory usage, often by a large margin. For instance, on the e-shop dataset with a minimum support of 10%, *SPOSR* successfully enumerates all rules within 270 s, whereas the competing approaches exceed the timeout of 1000 s and ERMiner runs out of memory (MO).

In summary, *SPOSR* requires less time and memory to enumerate all *POSRules* compared to the other algorithms. This superior performance can be attributed to the decomposition technique (see Sect. 4), which accelerates the

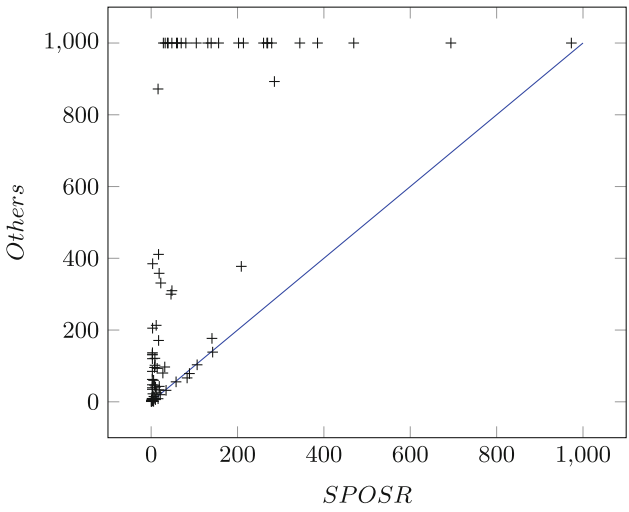


Fig. 1. Enumeration Time (in seconds): SPOSR vs Other Algorithms

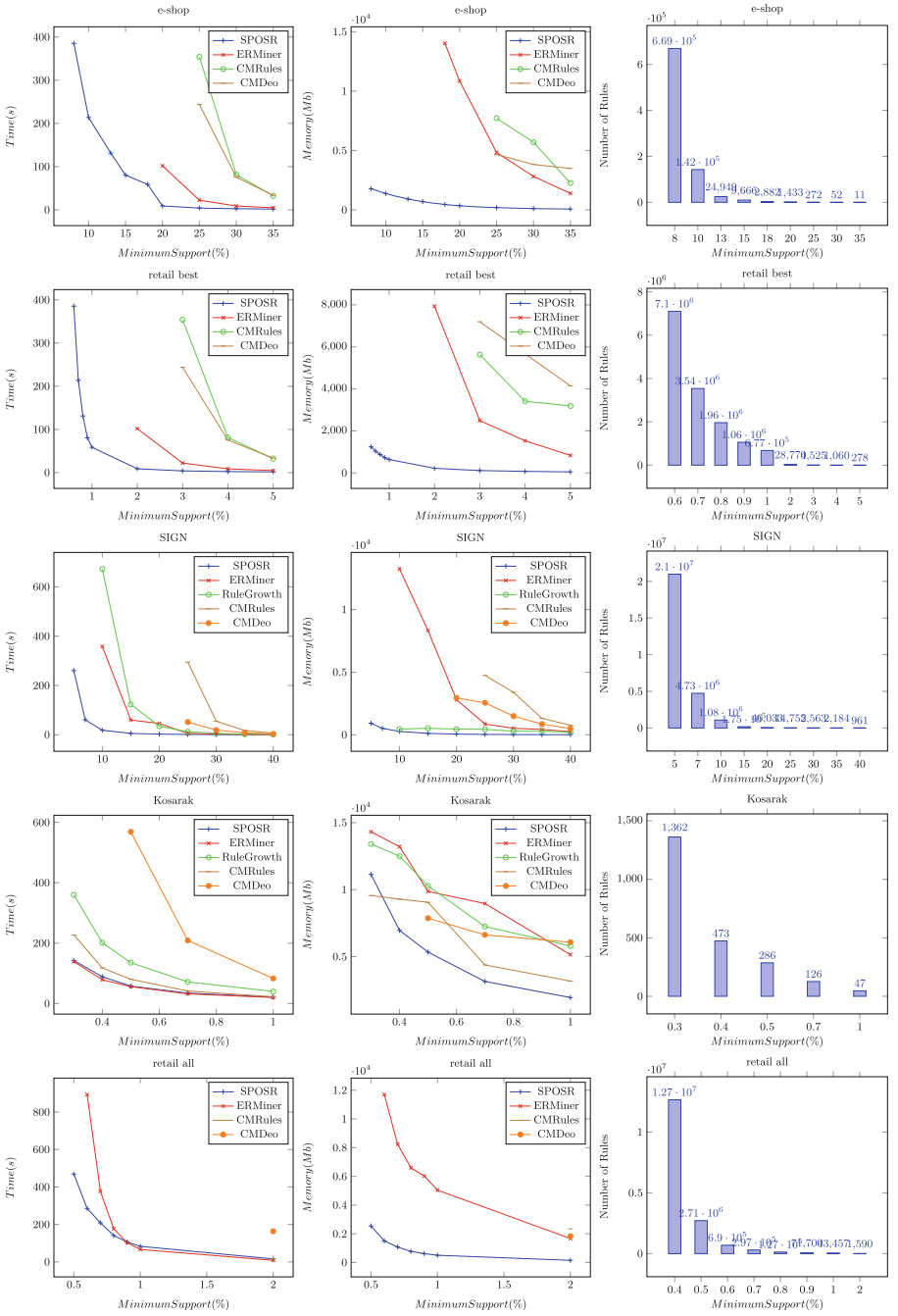


Fig. 2. Evaluation Results

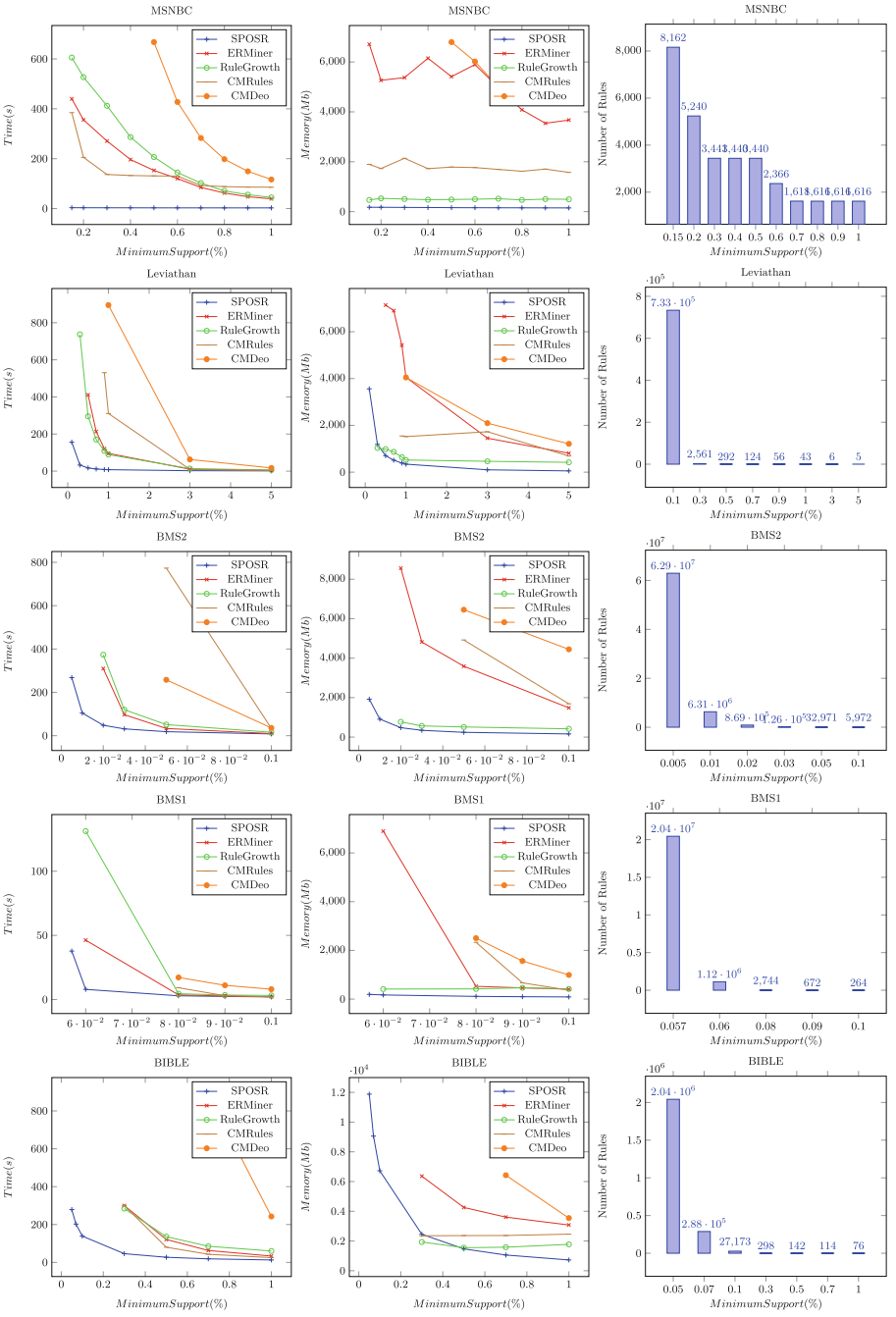


Fig. 3. Evaluation Results

enumeration process by pruning infrequent items from the search space, thereby reducing the size of the encoding.

6 Conclusion

In conclusion, this work addressed the problem of enumerating partially ordered sequential rules through a satisfiability-based framework. We introduced a SAT encoding that formalizes the enumeration of *POSRules* as a propositional logic formula, and demonstrated how a decomposition strategy can significantly enhance both the efficiency and scalability of the approach. Extensive experimental evaluations, along with a rigorous comparison against state-of-the-art methods, confirmed that the proposed approach consistently outperforms existing solutions in terms of execution time and memory usage across a wide range of datasets, including large-scale instances.

As a direction for future research, we aim to extend the proposed encoding to incorporate richer user-defined constraints and to enable the extraction of more compact and informative sets of sequential rules, such as minimal non-redundant rules.

References

1. Aoga, J.O.R., Guns, T., Schaus, P.: An efficient algorithm for mining frequent sequences with constraint programming. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2016), Lecture Notes in Computer Science, pp. 315–330. Springer (2016)
2. Bailleux, O., Boufkhad, Y., Roussel, O.: New encodings of pseudo-Boolean constraints into cnf. In: Proceedings of the SAT, pp. 181–194 (2009)
3. Boudane, A., Jabbour, S., Sais, L., Salhi, Y.: A sat-based approach for mining association rules. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pp. 2472–2478. IJCAI (2016)
4. Boudane, A., Ouzghara, R.: Compact sat-based encoding of mining minimal non-redundant association rules. In: Proceedings of the 2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–6. IEEE (2024)
5. Chen, Y., Fournier-Viger, P., Nouioua, F., Wu, Y.: Sequence prediction using partially-ordered episode rules. In: Proceedings of the IEEE ICDM Workshops, pp. 574–580. IEEE (2021)
6. Cook, S.A.: A short proof of the pigeonhole principle using extended resolution. SIGACT News **8**(4), 28–32 (1976)
7. De Raedt, L., Guns, T., Nijssen, S.: Constraint programming for itemset mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 204–212. ACM (2008)
8. Ezeife, C.I., Karlapalepu, H.: A survey of sequential pattern-based e-commerce recommendation systems. Algorithms **16**(10) (2023)

9. Fournier-Viger, P., Faghihi, U., Nkambou, R., Nguifo, E.M..: CMRULES: an efficient algorithm for mining sequential rules common to several sequences. In: Guesgen, H.W., Murray, R.C. (eds.) Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS), pp. 106–111. AAAI Press (2010)
10. Fournier-Viger, P., Faghihi, U., Nkambou, R., Nguifo,E.M.: CMRules: mining sequential rules common to several sequences. Knowl. Based Syst. **25**(1), 63–76 (2012)
11. Fournier-Viger, P., Gueniche, T., Zida, S., Tseng, V.S.: ERMiner: Sequential Rule Mining Using Equivalence Classes. In: Blockeel, H., van Leeuwen, M., Vinciotti, V. (eds.) IDA 2014. LNCS, vol. 8819, pp. 108–119. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12571-8_10
12. Fournier-Viger, P., Cheng-Wei, W., Tseng, V.S., Cao, L., Nkambou, R.: Mining partially-ordered sequential rules common to multiple sequences. IEEE Trans. Knowl. Data Eng. **27**(8), 2203–2216 (2015)
13. Hamadi, Y., Jabbour, S., Sais, L.: ManySAT: a parallel sat solver. J. Satisfiability Boolean Model. Comput. **6**(4), 245–262 (2009)
14. Izza, Y., Jabbour, S., Raddaoui, B., Boudane, A..: On the enumeration of association rules: a decomposition-based approach. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1265–1271. IJCAI (2020)
15. Jabbour, S., Lonlac, J., Sais, L., Salhi, Y..: Extending modern sat solvers for model enumeration. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI), pp. 803–810. IEEE (2014)
16. Jabbour, S., Mhadhbi, N., Raddaoui, B., Sais, L.: Sat-based models for overlapping community detection in networks. Computing **102**(5), 1275–1299 (2020)
17. Lee, J.M., Hauskrecht, M.: Neural clinical event sequence prediction through personalized online adaptive learning. In: Tucker, A., et al., (eds.) Artificial Intelligence in Medicine, pp. 175–186. Springer (2021)
18. Sinz, C.: Towards an optimal CNF encoding of Boolean cardinality constraints. In: Proceedings of the International Conference on Principles and Practice of Constraint Programming, pp. 827–831 (2005)
19. Tseitin, G.S.: On the complexity of derivations in the propositional calculus. Stud. Math. Math. Logic, 115–125 (1968)
20. Zaki, M.J.: SPADE: an efficient algorithm for mining frequent sequences. Mach. Learn. **42**(1–2), 31–60 (2001)

Detecting Real Software Faults Through Resource Consumption Anomalies: An Evaluation of the Tricorder Methodology

Caio Guimaraes Herrera[1(✉)], Simone R. S. Souza[1], Ricardo Ribeiro dos Santos[2], and Paulo Sergio Lopes de Souza[1]

[1] University of Sao Paulo (ICMC/USP), Sao Carlos, SP, Brazil
caio.herrera@outlook.com
[2] Federal University of Mato Grosso do Sul (UFMS), Campo Grande, MS, Brazil

Abstract. Software faults in production environments pose significant challenges to software reliability, as traditional testing approaches do not always account for non-deterministic or operational conditions. The Tricorder methodology addresses this gap by detecting software anomalies through the monitoring of resource consumption behavior, without requiring access to the application's internal logic. While previous studies validated the methodology against artificially inserted faults in controlled experiments, its effectiveness against real-world faults remains an open question. This paper presents the first evaluation of the Tricorder methodology on real software faults from the Defects4j dataset, using the TriAD (Tricorder-based Anomaly Detection) tool. A total of 55 faults were selected across seven open-source Java applications and eight fault classes. Results show that TriAD achieved a fault-detection rate of 33% and a control group correctness rate of 71%. A one-tailed Wilcoxon signed-rank test confirmed a statistically significant difference in the number of iterations required to detect actual faults versus control-group executions, producing an F1-score of 39.18%. These results demonstrate the potential of the Tricorder methodology to identify real software faults through resource-consumption anomalies, while highlighting challenges introduced by API constraints in the dataset that limited execution to unit tests. Future work will aim to evaluate TriAD in end-to-end application development scenarios.

Keywords: Software reliability · Anomaly detection · Software testing

1 Introduction

Software testing is essential for ensuring the quality of software systems, and remains one of the most widely used approaches in software engineering [1]. However, it is an expensive and time-consuming activity, as it requires detailed knowledge of program behavior and expected outputs to construct test cases. Moreover, traditional testing often fails to detect software faults caused by operational

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 146–161, 2027.
https://doi.org/10.1007/978-3-032-30494-0_10

conditions or non-deterministic aspects of production environments [13,16]. In this context, reliability engineering supports the fault detection processes by providing strategies for monitoring and continuously testing these applications, even in production environments [21].

The literature emphasizes runtime behavior modeling combined with statistical profiling to monitor systems and detect anomalies without access to internal code logic [3,6,7]. These approaches support the continuous evaluation of a system's reliability to identify deviations in the consumption of machine resources, such as RAM, CPU and I/O.

In this scenario, the **Tricorder** methodology [14] is presented as a complementary approach to regular software testing that aims to identify software faults through the monitoring of resource consumption, the definition of expected profiles, and the detection of anomalies. To evaluate its effectiveness, the methodology was implemented using a set of scripts and applied to selected applications in a laboratory environment, with artificial faults added to them. The **Tricorder** methodology detected the majority of these faults, demonstrating its effectiveness. However, it still lacks exposure to real faults in applications that were not tailored for conducting laboratory experiments. There is a need to determine whether **Tricorder** would remain effective at detecting real software faults in a realistic scenario.

This paper provides the first exposition of the **Tricorder** methodology applied to real software faults using the **Defects4j** dataset. We implemented the methodology as a reliability tool, **TriAD** (Tricorder-based Anomaly Detection), adapted to the dataset's API to execute unit tests for each selected application. An experiment was conducted to evaluate the methodology's effectiveness in this scenario and its behavior during control-group runs.

The remainder of this paper is organized as follows: Sect. 2 details how the **Tricorder** methodology works. Section 3 presents previous and related work to this paper, as well as explains how the methodology has been validated to date. Section 4 briefly explains how the **TriAD** tool was built, while Sect. 5 details how the **Defects4j** dataset was used, what were the needed adaptations to the **Tricorder** methodology, and how was the experiment set up. Finally, Sect. 6 presents and discusses the obtained results, and Sect. 7 concludes this paper, while discussing limitations and opportunities for future work.

2 The Tricorder Methodology

The **Tricorder** methodology [14] is an approach for detecting software anomalies by monitoring resource consumption of a system under test (SUT) as an indication of a possible fault in the code. The methodology relies on a baseline SUT version with no known faults as the reference behavior for resource consumption. Then, a new version of the SUT is introduced, and the new consumption behavior will be compared with the defined reference. This approach should not be confused with the Tricorder platform, proposed by Sadowski et al. [17], which performs static program analysis; the two share only the name.

Tricorder is intended to be used as a complementary approach to regular testing activities, as it does not rely on any structural or code information from the SUT, nor is it concerned about the correctness between input and output data. Moreover, **Tricorder**'s independence from code knowledge is the main factor enabling its use in production environments.

To define the expected resource consumption behavior for a particular SUT, **Tricorder** allows the definition of different workloads, which serve as one of the input parameters for executing the methodology. As a result, for each provided workload, the following steps will be executed:

1. Begin with the reference SUT version, **S**, and execute it **N** times.
2. Each execution will be monitored, and resource consumption data will be sampled to create a profile.
3. The profiles for the reference version are stored for future use.
4. Then, a new version of the SUT, called **S'**, will also be executed **N** times, and for each execution, a new profile is created.
5. After the generation of the profiles, a clustering step is executed. Starting from the full set of profiles in **S**, the clustering algorithm groups them in an unsupervised manner.
6. Profiles for **S'** are added one by one to the set, and the clustering process is executed again.
7. If any iteration yields a group containing only **S'** profiles, **Tricorder** is considered to have detected an anomaly.

The clustering step is conducted using the **DAMICORE** methodology [18], which enables unsupervised grouping of generic data. The goal in this step is to detect similarities between different executions of the SUT, where each group represents a different way it can behave. Groups composed of reference profiles are assumed to represent regular behavior, since the reference version serves as the baseline. On the other hand, groups containing only profiles from the new version are considered a new behavior; in this context, an anomaly.

To validate the effectiveness of the **Tricorder** methodology for detecting resource consumption anomalies, Python 2 scripts were developed to implement it and used in laboratory experiments. Two main scripts implement the methodology: a **monitoring** script that executes the SUT and samples the resource consumption into usage profiles, and an **analysis** script that runs clustering and issues a verdict on the existence of an anomaly. The **analysis** script uses **Damicorepy** [5], a Python implementation of **DAMICORE** that allows customization of its parameters.

Figures 1 and 2 illustrate the expected behavior for the **Tricorder** scripts when dealing with the reference version and an anomalous version of the SUT, respectively.

To support the resource consumption sampling, Montes et al. provided a utility tool within the **monitoring** script, called **Thermometer** [14], that relies on the *PSUtil* library to collect CPU usage, allocated bytes in RAM, and the

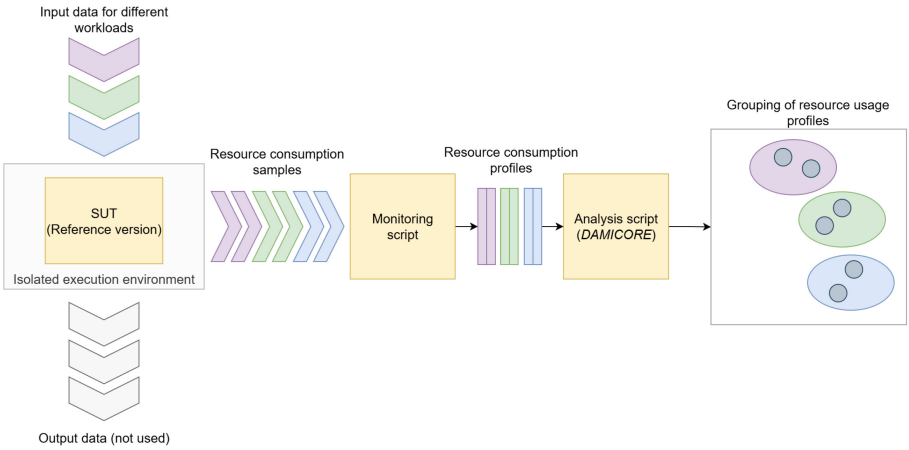


Fig. 1. Expected behavior for the **Tricorder** scripts, when running against the reference SUT version. Adapted from Herrera et al. [10].

number of I/O operations, which are used to build an execution profile. By default, metrics are sampled every 100 milliseconds. The authors reiterate that the **Tricorder** methodology allows the use of other metrics as well, and that the code can be extended to collect them.

Aiming to reduce false positives, Montes et al. also described and implemented a restrictive grouping strategy for the **analysis** script [14], in which an anomaly is reported only after being detected in two consecutive clustering iterations.

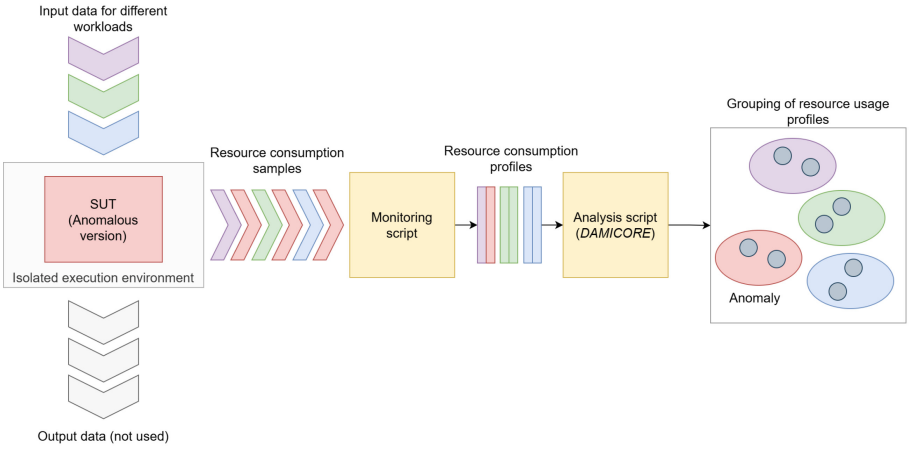


Fig. 2. Expected behavior for the **Tricorder** scripts, when running against an anomalous SUT version. Adapted from Herrera et al. [10].

3 Related Work

Based on the scripts presented in Sect. 2, a series of experiments was conducted to validate the effectiveness of the **Tricorder** methodology across different application domains and to investigate opportunities for performance improvement in its execution. Montes et al. started by instantiating **YATClient** and **YATServer** [14], a client-server application pair that simulates data stream traffic over the TCP protocol from the client to the server.

In this context, the **Tricorder** methodology monitored resource usage for the **YATServer** application. In the original application, faults were artificially inserted into the code, generating mutant versions. These new versions were used by **Tricorder** to detect anomalies in resource consumption behavior. Specifically, faults for seven distinct categories, inspired by real problems found in industrial environments [14], were inserted:

- **BCLEAN**: Excessive memory usage where allocated memory is not deallocated after use.
- **INFINITE**: An infinite loop causing unnecessary processing.
- **MONO**: A fault that limits parallelism, causing the application to run on a single thread.
- **SLEEP**: An unnecessary or excessively long pause in execution.
- **SWAP**: A fault in which the structure of the packages being sent does not have the same format as the expected structure to be received.
- **UNLOCK**: A synchronization fault in which a previously held mutex is not released.
- **NOBREAK**: A fault where a package of a specific type is processed as if it were from a different type.

Montes et al. showed that, despite some false positives, **Tricorder** detected all inserted mutants and achieved an F1-score of 91.4% in the experiment [14]. This is an important result for the methodology, as the **YATClient** and **YATServer** applications share characteristics of low-testability software [10], including processing data streams as their primary input.

In another study, Braga explored the effectiveness of the **Tricorder** methodology for different applications in the context of artificial intelligence and machine learning [4]. For these experiments, four distinct applications were selected as SUTs for the methodology:

- **BERT**[1]: Project that brings together different natural language processing (NLP) models, presenting a new and pre-trained model for language representation.
- **Chess-Alpha Zero**[2]: Project that uses reinforcement learning to reproduce a chess game, based on the **AlphaGo Zero** algorithm [22].

[1] https://github.com/google-research/bert.
[2] https://github.com/Zeta36/chess-alpha-zero.

- **Real-Time Voice Cloning**[3]: Project of a deep learning application that uses a text input and a five-second voice sample to generate audio.
- **Decision Tree**[4]: Standard machine learning project that uses the *scikit-learn* library to test different datasets [19].

In this multi-application scenario, Braga standardized the types of faults to be artificially inserted into the code. The generated mutants carried the following types of faults:

- **MODEL**: Machine learning configuration issues.
- **API**: Missing API call or calling an API at the wrong place.
- **TRAIN**: Missing training data.
- **PROCESS**: Issues when initializing an object or using a method.
- **LOGIC**: Logic errors in the code.
- **MEMORY**: Issues when using memory.
- **CONC**: Synchronization issues.

The **Tricorder** methodology was also effective in detecting the inserted anomalies in machine learning applications, as shown by Braga [4], where it correctly identified all mutants in **BERT** and **Chess-Alpha Zero** applications, while also detecting most of the faults in **Real-Time Voice Cloning** and **Decision Tree**, despite both applications presenting adverse conditions for the use of the methodology, where similar consumption behaviors were seen even for distinct workloads.

Durães also explored the effectiveness of the **Tricorder** methodology in detecting anomalies for different application domains [8]. Based on the same mutants defined by Montes et al. [14], Durães explored the following applications:

- **Crypto-Example**[5]: Cryptographic application, implemented as a proof of concept for RSA and AES methods, using the *OpenSSL* library.
- **JavaSort**[6]: Java project that implements the sorting algorithms *MergeSort* and *QuickSort*, in their parallel variants.
- **OBS Studio**[7]: Open-source video capture and transmitting tool, widely known by the community and popular among internet video creators.
- **Jet framework**[8]: A fluid simulation engine for computer graphic applications.

In their results, Durães highlighted the effectiveness of the **Tricorder** methodology in detecting anomalies for different domains, with an overall mean detection rate of 82% across all four applications [8], reiterating that the use

[3] https://github.com/CorentinJ/Real-Time-Voice-Cloning.
[4] https://doi.org/10.5281/zenodo.4723922.
[5] https://github.com/shanet/Crypto-Example.
[6] https://github.com/shaowdy/Multithreaded-Algorithms.
[7] https://obsproject.com/.
[8] https://github.com/doyubkim/fluid-engine-dev.

of **Tricorder** is not restricted to a particular class of applications. Moreover, a sample of the analyzed faults in the **OBS Studio** application came from real faults reported by developers, indicating that the methodology could also be effective in detecting real software anomalies rather than only artificial ones. These results serve as one of the motivations for this paper.

Finally, Fernandes et al. also explored the effectiveness of **Tricorder** for the **JavaSort** application, using parallel strategies to improve the methodology's performance [9]. As a result of this study, the authors showed that applying parallel strategies to the methodology can provide significant efficiency gains without compromising the quality of the results.

Despite all the studies providing concrete and positive evidence for the effectiveness of the **Tricorder** methodology in detecting software faults through resource consumption anomalies, there is still a need to explore the methodology's ability to detect anomalies in a real scenario, where faults are not artificially inserted into the experiment but are rather collected from the evolution of the software itself.

Beyond the **Tricorder** methodology, machine learning approaches have been applied to fault detection using the **Defects4j** dataset. For instance, Sekenov [20] evaluated multiple classifiers trained on static code metrics, achieving accuracy and F1-scores above 93% with Random Forest, and recall up to 96% with XGBoost. Similarly, Arab, Magel and Akour [2] framed fault localization as a multi-class classification problem, benchmarking deep learning, Random Forest, XGBoost, and LightGBM models, finding that test suite metrics consistently outperform static and dynamic metrics as fault predictors.

While these results are compelling, both approaches depend on access to the application's source code, limiting their applicability in scenarios where it is unavailable or proprietary. In contrast, the **Tricorder** methodology operates exclusively on runtime resource consumption, treating the application as a black box. This distinction makes **Tricorder** applicable in production environments where code inspection is not feasible—a trade-off that justifies a lower detection rate in exchange for broader applicability.

4 A Reliability Tool Based on the Tricorder Methodology

To expose **Tricorder** to real faults and facilitate its adoption, we implemented the methodology as a software reliability tool. The **TriAD** (Tricorder-based Anomaly Detection) tool [11] was developed based on the *RefTEST* architecture [15], enabling wider use of **Tricorder**. The adaptation of *RefTEST* to this new context is detailed in the work of Herrera et al. [10].

TriAD was developed as a web application divided into four modules: a front-end interface built with *Angular*; a back-end built in *Python 3*, *Flask* and *SQLite*; and two *Python 3* modules that replicate the original **monitoring** and **analysis** scripts. These technologies were selected for their simplicity, rapid prototyping

capabilities, and ability to reuse parts of the original code. This architecture is shown in Fig. 3, and the tool's source code is available on GitHub.[9].

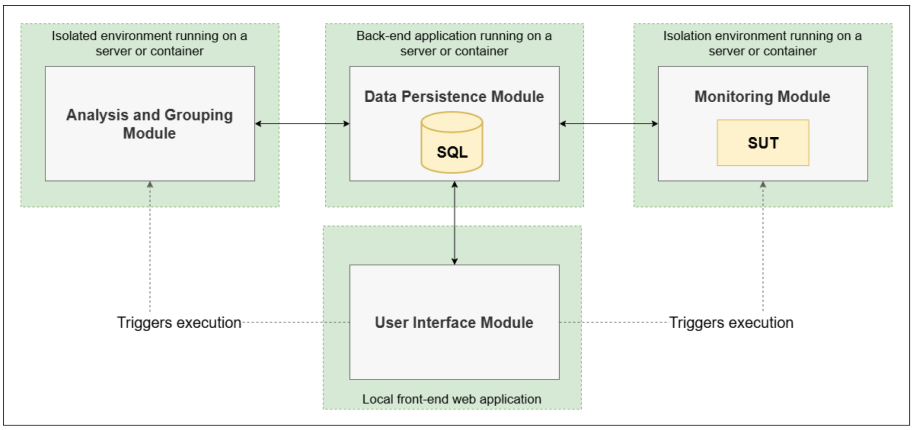


Fig. 3. Basic architecture used by the *TriAD* tool.

5 Experiment Design

To expose **Tricorder** (via **TriAD**) to real faults reported by users and developers, we selected **Defects4j** [12], a well-known dataset that summarizes faults, their descriptions, test cases that capture them, and the commits with fixes for several widely used open-source Java applications, including the main *Closure* compiler, *Apache* libraries *commons-lang* and *commons-math*, as well as *Mockito*, a framework for unit testing.

Spuy and Fischer classify 488 faults from **Defects4j** according to control and data flow representations [23]. This classification provides eight distinct classes of faults, listed below:

- **order**: Control flow fault, in which the order of instructions is incorrect;
- **jump**: Control flow fault, where an instruction A is followed by an instruction B when it should have been followed by an instruction C;
- **call**: Control flow fault, where a method is called when it should not be;
- **pred**: Control flow fault, in which a predicate condition (as an operator in a conditional or loop) is incorrect;
- **guard**: Control flow fault where a conditional or loop is missing to encapsulate a code block;
- **block**: Control flow fault, where there is a missing conditional statement or loop, and the code block that it encapsulates;

[9] https://github.com/caioherrera/triad-tricorder.

- **def**: Data flow fault, in which an incorrect value is written in a variable or the correct value is written in the incorrect variable;
- **use**: Data flow fault, where an incorrect variable is used for reading a value or an incorrect object is used for a method call.

Due to the large number of applications and faults in the dataset, a subset of **Defects4j** was chosen as the scope of this experiment. Seven out of seventeen applications were chosen; for each one, faults from each of the classes defined by Spuy and Fischer [23] were selected, with the exception of the **order** class for **Mockito**, since no fault of this class exists in that application. The selection of faults was conducted randomly, without prior knowledge of each one, with the only prioritization criteria being to choose faults with the least number of classes, in order to isolate the impact of each class as much as possible. The IDs for the selected faults can be seen in Table 1.

Table 1. IDs for the selected faults in the **Defects4j** dataset.

Project	order	jump	call	pred	guard	block	def	use
Chart	22	18	12	1	4	5	3	7
Closure	13	21	28	20	1	44	52	14
Jsoup	49	79	1	43	26	7	9	57
Lang	53	35	38	16	3	52	4	8
Math	86	1	35	15	28	44	2	34
Mockito	–	1	32	34	8	2	26	25
Time	7	15	26	19	3	24	4	16

To evaluate the effectiveness of the **Tricorder** methodology in this scenario, the following metrics were considered: anomaly detection rate and average number of iterations to detect an anomaly. We executed each application-fault combination 10 times and collected 30 profiles per execution. The reference version for each scenario was the application version with the fix to each fault, in order to provide the least number of changes between baseline and the test version.

To ensure that a high anomaly detection rate is reliable, it is also necessary to investigate the performance of **Tricorder** when evaluating the control group, i.e., a new execution of the reference version. The purpose of this evaluation is to check for false positives in the anomaly classification. It is expected that real anomalies require fewer iterations to be detected, while control group runs either produce no anomalies or require more iterations to produce one. For this analysis, we applied the **one-tailed Wilcoxon signed-rank** test to assess whether the median number of iterations differs significantly between fault detections and control group executions, as formalized in the hypotheses below:

- H_0: The number of iterations used in a control group execution is less than or equal to the number of necessary iterations to detect an anomaly.

– H_1: The number of iterations used in a control group execution is greater than the number of necessary iterations to detect an anomaly.

Despite the opportunity to explore distinct applications and fault classes in the context of the **Tricorder** methodology, using the **Defects4j** dataset poses additional challenges. Namely, it is not trivial to instantiate and execute end-to-end routines in the application. Instead, the dataset provides an API for executing individual test cases and sets of test cases for different commits. To work with these restrictions in the proposed scenario, we assume that the test cases provided by the developers exhibit behavior analogous to a conventional execution of the application and can therefore be used as an approximation of a regular application run. Still, a single unit test execution typically completes within milliseconds, well below the sampling interval established by **Tricorder**, which would yield profiles with too few data points to be meaningful. To address this, a script was created to repeatedly execute the test cases that capture each fault, collecting enough samples to generate a meaningful profile. While this makes the experiment viable, it introduces a threat to the validity of our results, that needs to be properly treated in future works.

6 Results and Discussion

The experiment was conducted in a virtual machine, hosted on a physical machine with the following specifications: a processor *Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz*, with 8 cores and 32GB DDR3 RAM, running *Ubuntu 24.04.2 LTS Server*. The VM also had 8 cores, 32GB RAM, and the same OS. During the experiment, only a single user accessed the machine at a time, providing an additional level of isolation.

Table 2 presents the anomaly detection rate for each combination of fault class and application, as listed in Table 1. Overall, the **Tricorder** methodology, implemented as the **TriAD** tool, detected 33% of the anomalies. Although the detection rate is lower when compared to previous work, this result is significant: it shows **Tricorder**'s potential to identify real faults by detecting anomalies in resource consumption. Another interpretation of this result is that a tester could have caught 33% of the anomalies before they reach end users.

An important factor affecting this result is the adaptation of **TriAD** to the **Defects4j** API. As described in Sect. 5, the dataset allows for the execution of test cases that capture the fault, not necessarily an end-to-end application flow. In a real scenario where **Tricorder** would be used alongside an application under development, this barrier would naturally be resolved, suggesting a more effective use of **TriAD** in detecting real faults.

Table 3 presents, within the same conditions, the performance of the **TriAD** tool when running against the control group. Overall, **TriAD** achieved a 71% correctness rate in this scenario. This result, similar to those obtained in previous studies, reiterates the method's expected behavior and indicates that the lower anomaly detection rate may be due to subtle anomalies that have a limited

Table 2. Anomaly detection rate for **TriAD**, when executed against the **Defects4j** dataset.

Project	order	jump	call	pred	guard	block	def	use	Average
Chart	30%	60%	40%	30%	50%	20%	40%	50%	40%
Closure	40%	40%	10%	50%	40%	50%	20%	10%	33%
Jsoup	20%	20%	30%	50%	30%	10%	20%	30%	26%
Lang	20%	40%	30%	30%	50%	50%	30%	30%	35%
Math	30%	30%	40%	50%	30%	40%	20%	30%	34%
Mockito	–	20%	30%	30%	30%	30%	30%	60%	33%
Time	40%	60%	20%	30%	40%	40%	20%	20%	34%
Average	30%	39%	29%	39%	39%	34%	26%	33%	**33%**

impact on resource consumption and are hard to detect. This suggests that under more favorable conditions, the effectiveness of **Tricorder** could be increased.

Table 3. Correctness rate of the control group for **TriAD**, when executed against the **Defects4j** dataset.

Project	order	jump	call	pred	guard	block	def	use	Average
Chart	70%	90%	70%	70%	50%	90%	40%	50%	66%
Closure	90%	60%	80%	60%	80%	90%	80%	80%	78%
Jsoup	100%	70%	90%	70%	60%	60%	80%	80%	76%
Lang	50%	60%	80%	40%	70%	80%	80%	60%	65%
Math	80%	70%	90%	60%	60%	80%	90%	80%	76%
Mockito	-	70%	70%	50%	80%	70%	70%	40%	64%
Time	80%	70%	90%	60%	90%	50%	70%	70%	73%
Average	78%	70%	81%	59%	70%	74%	73%	66%	**71%**

Table 4 shows the average number of iterations needed to detect an actual fault. In contrast, Table 5 shows the average number of iterations used for control group executions. In an ideal scenario, we expect the former to be as few as possible, with the latter as close as possible to the maximum number of iterations. In our experiment, fault detection required an average of 26.5 iterations, while control group runs averaged 27.3 iterations.

Considering $\alpha = 0.05$, we can use the **one-tailed Wilcoxon signed-rank** on the data presented in Tables 4 and 5, obtaining an effective sample size of $N_{effective} = 53$. Furthermore, we get the value of $W_{crit} = 902$, with $W+ = 998.5$. From a normal approximation to a binomial distribution ($N_{effective} > 25$), we obtain $p \approx 0.006 < \alpha$, allowing us to reject the null hypothesis (H_0) in favor of the alternative hypothesis (H_1) and to assert that there is statistically

Table 4. Average number of iterations needed to detect a fault through **TriAD**, when executed against the **Defects4j** dataset.

Project	order	jump	call	pred	guard	block	def	use	Average
Chart	26.0	26.1	25.6	25.0	22.9	27.8	23.2	23.8	25.1
Closure	26.3	29.0	29.5	25.0	27.4	23.0	28.1	29.7	27.3
Jsoup	26.0	26.7	28.3	22.7	26.6	29.5	29.7	25.2	26.8
Lang	26.1	23.6	28.2	29.5	24.8	24.6	26.9	26.3	26.3
Math	28.1	27.0	26.0	22.6	29.6	26.2	27.5	26.4	26.7
Mockito	-	28.7	29.0	25.8	27.5	25.6	25.9	23.1	26.5
Time	28.2	24.3	27.2	29.1	25.6	26.9	27.0	29.3	27.2
Average	26.8	26.5	27.7	25.7	26.3	26.2	26.9	26.3	**26.5**

Table 5. Average number of iterations used for control group executions for **TriAD**, when executed against the **Defects4j** dataset.

Project	order	jump	call	pred	guard	block	def	use	Average
Chart	28.3	28.9	26.1	26.1	27.2	29.4	24.4	25.6	27.0
Closure	29.1	27.8	28.2	26.2	27.2	30.0	25.9	28.3	27.8
Jsoup	30.0	26.8	29.4	28.3	26.3	28.1	29.1	28.2	28.3
Lang	23.8	25.7	27.7	25.3	25.1	27.1	26.9	28.3	26.2
Math	27.2	27.0	28.9	26.4	26.5	28.7	29.7	28.1	27.8
Mockito	-	27.9	26.3	26.5	27.7	27.1	27.4	22.0	26.4
Time	27.4	26.3	29.7	27.7	29.2	26.1	27.4	27.8	27.7
Average	27.6	27.2	28.0	26.6	27.0	28.1	27.3	26.9	**27.3**

significant evidence suggesting that the median of iterations used in control group executions is greater than the median of iterations needed to detect actual faults. This is an important result, pointing to the effectiveness of the **Tricorder** methodology, implemented in the **TriAD** tool, in detecting real software faults.

It should be noted, however, that the practical gap between the two groups is narrow. Repeated invocations of short tests tend to produce resource profiles with low variance, reducing the separation between anomalous and normal behavior. In an end-to-end execution scenario, a wider and more reliable gap would be expected. This limitation is further reflected by the confusion matrix for **TriAD**, shown in Table 6. When evaluated against the selected faults, the **Tricorder** implementation achieved a precision of 52.44%, a recall of 31.27%, and an F1-score of 39.18%, indicating clear opportunities for improvement in future works.

Table 6. Confusion matrix for **TriAD**, when executed against the **Defects4j** dataset.

	Predicted Anomalous	Predicted Normal
Actual Anomalous	172 (True Positive)	378 (False Negative)
Actual Normal	156 (False Positive)	394 (True Negative)

7 Conclusion and Future Work

This work demonstrated the effectiveness of the **Tricorder** methodology, implemented as the **TriAD** reliability tool, when applied to real faults from well-known applications in the **Defects4j** dataset. We selected 55 faults from eight classes and seven distinct applications for investigation using the methodology. In this scenario, despite the need for adjustments and adaptations to the tool to work with the dataset-provided API, **Tricorder** achieved a fault detection rate of 33% while maintaining a correctness rate of 71% in control group executions. These results indicate the effectiveness of the **Tricorder** methodology, even though it still needs improvement, as it would be able to detect 33% of anomalies before the software reaches its final users in a production environment.

Beyond that, the **TriAD** tool needed, on average, 26.5 iterations to successfully detect an anomaly, whereas it used 27.3 iterations on average for control group executions. Through the **one-tailed Wilcoxon signed-rank** statistical test, we were able to confirm that there is a statistically significant difference between these two groups, allowing us to state that the median of iterations needed to detect a fault is smaller than the median of iterations used by the control groups, despite the small difference between them.

We obtained an F1-score of 39.18%, indicating room for improvement. This result reflects both the concessions made to enable the experiment and the inherent difficulty of detecting real faults, which are not designed to maximally impact resource consumption, unlike artificially inserted mutants. It is worth noting that **Tricorder** operates under fundamentally different constraints than white-box approaches, since it does not rely on access to the application's source code. Even so, a 33% detection rate means one in three real faults could be caught before reaching production, with no code instrumentation required—a meaningful result for scenarios where source-level analysis is infeasible.

Some limitations and threats to validity were identified during the experiment. Mainly, the necessary adaptations that were conducted to enable the use of the **Defects4j** dataset. While it provided a diverse set of applications and fault classes, it constrained the experiment to unit test execution - an approximation that, as discussed in Sect. 5, might have reduced the fault's impact on resource consumption. End-to-end runs could amplify anomalous behavior across the full execution path, potentially yielding better results, which requires further investigation.

Future work could address these challenges, while expanding the scope of the experiment to the full **Defects4j** dataset, enabling deeper analysis, including the investigation of fault classes that are more or less susceptible to resource consumption anomalies, and whether detection rates scale with application size or fault complexity. Additionally, the **TriAD** tool could be used in an end-to-end application development scenario, monitoring actual application runs rather than individual test executions.

Acknowledgments. The authors would like to thank CAPES (PROEX), CNPq (306719/2025-8, 311207/2025-1, 406941/2025-4), and FAPESP (2013/07375-0, 2019/06937-0, 2019/26702-8) for their financial support.

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

References

1. Ammann, P., Offutt, J.: Introduction to Software Testing. Cambridge University Press, Cambridge (2016)
2. Arab, I., Magel, K., Akour, M.: Evaluating the predictive power of software metrics for fault localization. Computers **14**(6) (2025). https://doi.org/10.3390/computers14060222
3. Baah, G.K., Gray, A., Harrold, M.J.: On-line anomaly detection of deployed software: a statistical machine learning approach. In: Proceedings of the 3rd International Workshop on Software Quality Assurance, pp. 70–77. Association for Computing Machinery (2006). https://doi.org/10.1145/1188895.1188911
4. Braga, D.: (in portuguese) Avaliação do uso de agrupamento de dados de desempenho para apoiar o teste de software no domínio de aprendizagem de máquina. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos (2022). https://doi.org/10.11606/D.55.2022.tde-09112022-160513
5. Cesar, B.K.M.: (in portuguese) Estudo e extensão da metodologia DAMICORE para tarefas de classificação. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil (2016). https://doi.org/10.11606/D.55.2016.tde-16112016-170837
6. Cherkasova, L., Ozonat, K., Mi, N., Symons, J., Smirni, E.: Automated anomaly detection and performance modeling of enterprise applications. ACM Trans. Comput. Syst. **27**(3) (2009). https://doi.org/10.1145/1629087.1629089
7. Dean, D.J., Nguyen, H., Gu, X.: UBL: unsupervised behavior learning for predicting performance anomalies in virtualized cloud systems. In: Proceedings of the 9th International Conference on Autonomic Computing. ICAC '12, New York, NY, USA, pp. 191–200. Association for Computing Machinery (2012). https://doi.org/10.1145/2371536.2371572

8. Durães, T.d.J.O.: (in portuguese) Avaliação do uso de perfis de desempenho aplicados no teste e software em diferentes domínios de aplicação. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos (2022). https://doi.org/10.11606/D.55.2022.tde-05042023-144747
9. Fernandes, M.L., Souza, P.S.L.D.: Parallel implementation for anomaly detection in methodologies based on consumption profiles. In: 51st Latin American Informatics Conference (CLEI) (2025)
10. Guimaraes Herrera, C., Silva Montes, V., S. Souza, S.R., Ribeiro dos Santos, R., Lopes de Souza, P.S.: An architecture for a reliability tool applied to distributed systems. In: Computational Science and Its Applications – ICCSA 2025, pp. 332–348 (2025)
11. Herrera, C.G.a., Souza, S.d.R.S.d., Santos, R., Souza, P.S.L.d.: TriAD: a Tricorder-based tool for detecting anomalies in software. In: Proceedings of the 19th IEEE International Conference on Software Testing, Verification and Validation (ICST) – Testing Tools and Data Showcase Track (2026, to appear)
12. Just, R., Jalali, D., Ernst, M.D.: Defects4j: a database of existing faults to enable controlled testing studies for java programs. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis. ISSTA 2014, New York, NY, USA, pp. 437–440. Association for Computing Machinery (2014). https://doi.org/10.1145/2610384.2628055
13. Liao, L., et al.: Using black-box performance models to detect performance regressions under varying workloads: an empirical study. Empirical Softw. Engg. **25**(5), 4130–4160 (2020). https://doi.org/10.1007/s10664-020-09866-z
14. Montes, V.S., Herrera, C.G., Durães, T.J.O., Delbem, A.C.B., de Souza, S.R.S.L., Souza, P.S.L.d.: Software reliability forecasting using anomalies in profiles of resource usage. In: 2025 IEEE 36th International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 189–196 (2025). https://doi.org/10.1109/ISSREW67781.2025.00071
15. Nakagawa, E., Simão, A., Ferrari, F., Maldonado, J.: Towards a reference architecture for software testing tools, pp. 157–162 (2007)
16. Patel, K., Hierons, R.: A mapping study on testing non-testable systems. Software Qual. J. **26** (2018). https://doi.org/10.1007/s11219-017-9392-4
17. Sadowski, C., Van Gogh, J., Jaspan, C., Soderberg, E., Winter, C.: Tricorder: Building a program analysis ecosystem. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, pp. 598–608 (2015). https://doi.org/10.1109/ICSE.2015.76
18. Sanches, A., Cardoso, J.M., Delbem, A.C.: Identifying merge-beneficial software kernels for hardware implementation. In: 2011 International Conference on Reconfigurable Computing and FPGAs, pp. 74–79 (2011). https://doi.org/10.1109/ReConFig.2011.51
19. Santos, S.A., Silveira, B., Durelli, V., Durelli, R., Souza, S., Delamaro, M.: On using decision tree coverage criteria fortesting machine learning models. In: Proceedings of the 6th Brazilian Symposium on Systematic and Automated Software Testing. SAST '21, New York, NY, USA, pp. 1–9. Association for Computing Machinery (2021). https://doi.org/10.1145/3482909.3482911
20. Sekenov, G.: Machine learning approaches for predicting defective java code using software metrics. Int. J. Digit. **1**, 1–9 (2026). https://doi.org/10.20508/9j0jxz49
21. Shieh, S.W., et al.: Reliability engineering in a time of rapidly converging technologies. IEEE Trans. Reliab. **73**(1), 73–82 (2024). https://doi.org/10.1109/TR.2024.3355905

22. Silver, D., et al.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm (2017)
23. van der Spuy, A., Fischer, B.: An anatomy of 488 faults from defects4j based on the control- and data-flow graph representations of programs. In: Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering. EASE '25, New York, NY, USA, pp. 622–627. Association for Computing Machinery (2025). https://doi.org/10.1145/3756681.3757025

Enhancing the Trustworthiness of Deep Ensemble via Explainability in Medical Imaging

Gbègninougbo Aurel Davy Tchokponhoue[1] and Ali Idri[2(✉)]

[1] Faculty of Medical Science, UM6P Hospitals, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco
aurel.tchokponhoue@um6p.ma

[2] Data and Software Sciences, ENSIAS, Mohammed V University, Rabat, Morocco
ali.idri@ensias.um5.ac.ma

Abstract. Deep Ensembles are among the best methods for uncertainty quantification in machine learning, especially in high-stakes areas like medical imaging. However, explaining how they work is still a challenge. To address this, we present two algorithms for their explainability: *Ensemble Conditioned Class Activation Map (ECAM)* and *Selective Ensemble Class Activation Map (SECAM)*. ECAM uses adaptive, sample-specific weighting based on predictive entropy to combine explainability maps that match ensemble predictions, which improves interpretability. SECAM uses the ADCC metric to choose the best explainability map from ensemble members whose predictions are closest to the output of the ensemble. This provides a model-specific view. We tested the suggested methods on a breast cancer ultrasound dataset (BUSI), which includes ground truth masks for localization assessment. The results show that ECAM and SECAM highlight important clinical areas and align well with the ground truth masks. These findings suggest that the proposed methods can improve the transparency of deep ensemble models and contribute to more trustworthy AI systems for medical diagnosis.

Keywords: Explainability · Deep ensemble · Uncertainty quantification · Trustworthy AI

1 Introduction

Deep learning (DL) has become a game-changing tool in medical diagnostics, showing unprecedented accuracy in various clinical tasks like imaging-based disease detection, predicting outcomes, and treatment planning [18]. However, the deployment of DL models in real-world clinical contexts remains limited due to concerns about their reliability, interpretability, and trustworthiness [7]. In high-stakes fields like healthcare, uncertainty quantification (UQ) is very important

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 162–176, 2027.
https://doi.org/10.1007/978-3-032-30494-0_11

for making reliable models because it shows how sure the model's predictions are.

In recent years, Monte Carlo Dropout, Bayesian Neural Networks, and Deep Ensembles are some of the UQ methods that have been investigated in the healthcare domain [7]. Among these, Deep Ensemble models have become the state-of-the-art, providing great results in both predicting accuracy and estimating uncertainty [2,6]. However, the explainability of these models continues to pose a challenge and remains largely unexamined in UQ studies for imaging applications, resulting in a significant gap.

In the wider area of ensemble learning applied to imaging, attempts to improve explainability often centered on combining interpretation maps created by individual ensemble members [4,11,15,19]. These methods were inspired by research that combined maps from different interpretability techniques for a single network [5,9,13,16]. They used techniques like averaging, summation, median calculation, or attention mechanisms. These strategies showed that they could produce good explainability maps for the prediction of a single model, setting the stage for further research in ensemble learning [13].

Building on this foundation, some studies on ensemble methods used similar aggregation strategies while others extended them to new ones [21]. Specifically, they introduced weighted aggregation schemes, where the weights assigned to individual model maps are based on the validation performance of each ensemble member [11]. Although these approaches usually achieve acceptable results, they have some limitations. First, they ignore the sample-specific variability of each model performance, assuming that each model works the same way across the dataset. However, a model member that performs well overall may struggle with certain cases, leading to unreliable or misleading explainability maps for those instances. Second, these methods produced an explainability map that is independent of the final prediction of the ensemble. This is an important drawback because the predictions from individual models can sometimes contradict the ensemble's overall decision. As a result, the combined maps may include information that does not match the ensemble's logic, adding noise and reducing the overall reliability of the explainability.

Class Activation Mapping (CAM) techniques [22] are widely used for explainability in DL, as they enable visualization of image regions that most influence a model's predictions. This paper describes two novel explainability algorithms developed for Deep Ensemble models, both leveraging CAM techniques. The first algorithm, Ensemble Conditioned Class Activation Map (ECAM), combines the explainability maps of individual ensemble members and weights their contributions according to predictive uncertainty, rather than using uniform or validation-based weighting. This method conditions the explainability on the ensemble's collective decision, producing a coherent and sample-specific representation. The second algorithm, Selective Ensemble Class Activation Map (SECAM), uses a selective approach, identifying and using only the most relevant explainability maps from the individual models that match the ensemble's overall prediction, reducing noise and inconsistencies. This strategy is inspired

by the use of surrogate models, which are commonly used to interpret complex black-box models.

ECAM and SECAM address an important gap in the explainability of Deep Ensemble models, improving the transparency of this UQ method. We show the effectiveness of these algorithms in breast cancer screening using ultrasound data. By comparing the explainability maps generated by the ensemble model to the ground truth masks, we evaluate their ability to capture clinically relevant features and accurately reflect the ensemble's decision-making process.

In Sect. 2, we provide an overview of related work on explainability in deep learning, with a focus on imaging applications. Section 3 offers background information on the deep ensemble method and the Class Activation Mapping (CAM) strategy employed in this study. In Sect. 4, we introduce two novel Deep ensemble explainability algorithms. Section 5 presents the experimental results, including the generated maps produced by the algorithms. Finally, Sect. 6 discusses the limitations of the study, and Sect. 7 concludes the paper, offering suggestions for future research directions.

2 Explainability of Deep Learning-Based Imaging Applications

Explainability methods are crucial in deep learning, especially for image-based tasks, where understanding the decision-making process of black-box convolutional model is vital for sensitive domains like healthcare and autonomous systems. Local Interpretable Model-Agnostic Explanations (LIME) addresses this by creating local surrogate models to explain predictions, identifying important features through perturbations while balancing interpretability using a complexity penalty [12]. Similarly, SHapley Additive exPlanations (SHAP), based on game theory, computes Shapley values to attribute feature contributions to predictions, providing detailed insights into individual and interactive feature effects [8]. CAM techniques identify the important regions of the image that drive the model decision by combining weights from the final fully connected layer with activations from the last convolutional layer, generating class-specific heatmaps for CNNs [22]. Grad-CAM extends CAM by using backpropagated gradients with feature maps to highlight the regions most relevant to specific predictions and improving visual explanations [14]. Enhanced variants like Grad-CAM++, XGradCAM, and ScoreCAM give stronger and visually refined explanations. In the context of ensemble models, Rajani and Raymond introduced Weighted Average Ensemble Explanation and Penalized Weighted Average Ensemble Explanation for Visual Question Answering (VQA), where they used validation score as a global weight for model aggregation [11]. In contrast, we propose a new approach that uses sample-level weights derived from the predictive entropy of each ensemble member for the particular prediction. This led to adaptive explanations that moved beyond naive aggregation to better reflect the behavior of the Deep Ensemble method.

3 Background

3.1 Deep Ensemble

Deep ensembles [6] have become a strong method for estimating predictive uncertainty in DL. Unlike BNN, which need complex changes and are heavy on computation, Deep Ensembles provide a simpler and more scalable option. The idea is to train multiple neural networks independently, with each network initialized with different weights, and then combine their predictions to assess uncertainty. This method captures the differences in model outputs that result from different weight settings. It effectively measures both epistemic and aleatoric uncertainty.

Let $p(y \mid \mathbf{x})$ represent the predicted posterior distribution over the classes for a given input $\mathbf{x}$. In the case of an ensemble comprising M models, the predictive distribution is calculated by averaging the outputs of the individual models:

$$p(y \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_{\theta_m}(y \mid \mathbf{x}), \tag{1}$$

where θ_m denotes the parameters of the m-th model.

Epistemic uncertainty is quantified by calculating the variance of ensemble predictions:

$$\sigma^2\big[p(y \mid \mathbf{x})\big] = \frac{1}{M} \sum_{m=1}^{M} \left(p_{\theta_m}(y \mid \mathbf{x}) - p(y \mid \mathbf{x})\right)^2. \tag{2}$$

The aleatoric uncertainty is measured using the entropy of the predictive distribution:

$$H\big[p(y \mid \mathbf{x})\big] = -\sum_{y \in \mathcal{Y}} p(y \mid \mathbf{x}) \log p(y \mid \mathbf{x}). \tag{3}$$

3.2 Score-CAM

Score-CAM [20] is a method for explaining the decisions of convolutional neural networks (CNNs). It highlights the areas of an image that have a big impact on the model's predictions. Unlike methods that rely on gradients, Score-CAM avoids using gradients, which makes the explanations less sensitive to gradient noise and improves their robustness. To start, the method gets activation maps from a convolutional layer that has been chosen. After that, each activation map is resized to fit the size of the input image and normalized to make a mask. This mask is applied to the original image, and the modified image is fed back into the network to observe changes in the target class score. The increase in the target class score from the masked image acts as a weight for the corresponding activation map. The final class activation map is created by calculating a weighted sum of these activation maps. This method highlights the most important areas

of the input image. Score-CAM has shown better performance in both recognition and localization tasks compared to earlier methods. It offers clearer visual explanations without needing gradient information.

Given a convolutional layer l of a model f and a target class c, the Score-CAM localization map is defined as:

$$L^{c}_{\text{Score-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}_{l}\right), \tag{4}$$

where A^{k}_{l} denotes the k-th activation map at layer l, and α^{c}_{k} is the importance weight defined as:

$$\alpha^{c}_{k} = \text{softmax}_{k}\left(Y^{(c)}(M_k) - Y^{(c)}(\mathbf{X}_b)\right), \tag{5}$$

with $Y^{(c)}(\cdot)$ denoting the pre-softmax score for class c, $\mathbf{X}_b$ a baseline input, and M_k defined as:

$$M_k = \left(\frac{U(A^{k}_{l}) - \min_m U(A^{m}_{l})}{\max_m U(A^{m}_{l}) - \min_m U(A^{m}_{l})}\right) \odot \mathbf{X}_b, \tag{6}$$

where $U(\cdot)$ denotes upsampling and $\odot$ element-wise multiplication.

3.3 Average DCC (ADCC)

The Average DCC (ADCC) explainability map evaluation metric [10], was employed to develop our second algorithm, which adopts a selective CAM strategy. This strategy identifies the explainability map that best matches the output of the Deep Ensemble model. ADCC combines three key scores: Maximum Coherency, Minimum Complexity, and Minimum Average Drop, using a harmonic mean, as shown in (7):

$$\text{ADCC}(\mathbf{x}) = 3\left(\frac{1}{\text{Coherency}(\mathbf{x})} + \frac{1}{1 - \text{Complexity}(\mathbf{x})} + \frac{1}{1 - \text{AverageDrop}(\mathbf{x})}\right)^{-1}. \tag{7}$$

The Coherency score component makes sure that the CAM captures all relevant features that explain a prediction. It also removes any irrelevant or redundant features in a clear and logical manner, as defined in (8).

$$\text{Coherency}(\mathbf{x}) = \frac{\text{Cov}\big(\text{CAM}_c(\mathbf{x}) \odot \text{CAM}_c(\mathbf{x}), \text{CAM}_c(\mathbf{x})\big)}{\sigma_{\text{CAM}_c(\mathbf{x}) \odot \text{CAM}_c(\mathbf{x})}\, \sigma_{\text{CAM}_c(\mathbf{x})}}. \tag{8}$$

The complexity score promotes simplicity by requiring the CAM to show the smallest set of pixels that sufficiently explains the prediction, as defined in (9).

$$\text{Complexity}(\mathbf{x}) = \|\text{CAM}_c(\mathbf{x})\|_1. \tag{9}$$

The Average Drop part means that a good explanation map should make the confidence drop as little as possible when the masked input, based on the

explanation map, is used instead of the original image. A smaller drop shows that the explanation map retains the most important information for the prediction.

Recently, new metrics have come up to improve the evaluation of these maps, especially by using segmentation-based ground truth annotations. For example, [17] proposed the Explanation Significance Assessment (ESA) and its WESA extension.These methods use pixel-level spatial analysis and distance measurements to evaluate how well explanation maps match with annotated regions. In contrast, the ADCC metric we use in our study does not depend on ground-truth annotations. This makes it suitable for fully weakly-supervised settings without needing extra labeling.

4 Novel Deep Ensemble Explainability Algorithms

In this section, we present two new explainability algorithms, ECAM and SECAM, created for Deep Ensemble models. These algorithms make the model predictions easier to understand by providing clear explanations that reflect the decision-making process of the ensemble. We give a detailed description of each algorithm and explain their basic principles.

4.1 Ensemble Conditioned Class Activation Map (ECAM)

Algorithm 1 introduces *Ensemble Conditioned Class Activation Map (ECAM)*, a novel method to build an explainability map for deep ensemble by integrating the outputs of the individual models in the deep ensemble. Unlike existing approaches that combine class activation maps (CAMs) using uniform averaging or validation-based weighting, ECAM leverages *uncertainty-aware metrics* derived from the *normalized inverse entropy* of each model's predictive distribution.

Moreover, while traditional methods compute model-specific CAMs independently, Algorithm 1 ensures that the CAM of each member of the ensemble M_t is conditioned on the global prediction of the ensemble $\mathbf{y}_{\text{ens}}$. This design matches each explainability map with the group's behavior of the ensemble.

Given an input image $\mathbf{X}$, each model M_t, for $t = 1, \ldots, T$, produces a CAM, noted as $\text{CAM}_t(\mathbf{X})$, based on $\mathbf{y}_{\text{ens}}$. The reliability weight for each model is calculated using its predictive entropy $H(\mathbf{p}_t)$ as:

$$w_t = 1 - \frac{H(\mathbf{p}_t)}{\log C}, \tag{10}$$

where $\mathbf{p}_t \in \mathbb{R}^C$ is the predicted probability distribution of the model M_t.$H(\mathbf{p}_t)$ is the predicted entropy and C is the number of output classes.

The weights $\{w_t\}_{t=1}^T$ are normalized so that $\sum_{t=1}^T w_t = 1$. The final ECAM output is computed as the weighted sum of individual CAMs:

$$\text{ECAM}(\mathbf{X}) = \sum_{t=1}^{T} w_t \cdot \text{CAM}_t(\mathbf{X}). \tag{11}$$

This formulation makes sure that the resulting explanation focuses on the more reliable members of the group. This improves both the interpretability and how well it matches the overall decision of the group.

Algorithm 1. Ensemble Conditioned Class Activation Map (ECAM)

Require: Input image $\mathbf{X}$; Ensemble models $\{M_1, \ldots, M_T\}$; Number of classes C; Ensemble prediction $\mathbf{y}_{\text{ens}}$
Ensure: Explanation map ECAM($\mathbf{X}$)
Initialize empty lists $\mathcal{W} \leftarrow [\,]$ and $\mathcal{C} \leftarrow [\,]$
for $t = 1$ to T **do**
  $\mathbf{p}_t \leftarrow \text{softmax}(M_t(\mathbf{X}))$
  $\text{CAM}_t(\mathbf{X}) \leftarrow \text{ScoreCAM}(M_t, \mathbf{X}, \mathbf{y}_{\text{ens}})$
  $H_t \leftarrow -\sum_{c=1}^{C} \mathbf{p}_t[c] \log \mathbf{p}_t[c]$
  $\tilde{w}_t \leftarrow 1 - \frac{H_t}{\log C}$
  Append $\tilde{w}_t$ to $\mathcal{W}$
  Append $\text{CAM}_t(\mathbf{X})$ to $\mathcal{C}$
end for
for $t = 1$ to T **do**
  $w_t \leftarrow \frac{\mathcal{W}[t]}{\sum_{j=1}^{T} \mathcal{W}[j]}$
end for
Compute $\text{ECAM}(\mathbf{X}) \leftarrow \sum_{t=1}^{T} w_t \cdot \mathcal{C}[t]$
return ECAM($\mathbf{X}$)

4.2 Selective Ensemble Class Activation Map (SECAM)

Algorithm 2, *Selective Ensemble Class Activation Map (SECAM)*, presents a focused method for creating explainability maps. It picks only the most relevant ensemble models, specifically those whose predictions match the overall output of the ensemble. Unlike current methods that combine all ensemble member CAMs or calculate them without considering the ensemble's output, SECAM filters ensemble members. This process ensures that the final explanation comes from the most trustworthy and semantically aligned model.

Given an input image $\mathbf{X}$, each model M_t in the ensemble is checked for eligibility based on whether its prediction $\mathbf{y}_t$ matches the ensemble's global prediction $\mathbf{y}_{\text{ens}}$. The set of eligible models is defined as:

$$\mathcal{M}_{\text{eligible}} = \{M_t \mid t \in \{1, \ldots, T\},\ \mathbf{y}_t = \mathbf{y}_{\text{ens}}\},$$

For each eligible model $M_t \in \mathcal{M}_{\text{eligible}}$, a class activation map $\text{CAM}_t(\mathbf{X})$ and an associated alignment score $\text{ADCC}_t(\mathbf{X})$ are computed. The alignment score $\text{ADCC}_t(\mathbf{X})$ serves as a measure to assess the quality of the explanation provided by CAM_t.

The final SECAM explanation comes from the CAM linked to the eligible model with the highest alignment score. By using both an eligibility criterion and the alignment score, SECAM makes sure that the final explanation comes from

the most relevant and reliable model in the group. This improves robustness and prevents noisy contributions from misaligned predictions.

Algorithm 2. Selective Ensemble Class Activation Map (SECAM)

Require: Input image $\mathbf{X}$; Ensemble models $\{M_1, \ldots, M_T\}$; Ensemble prediction $\mathbf{y}_{\text{ens}}$
Ensure: Explanation map SECAM($\mathbf{X}$)
Initialize $s^* \leftarrow -\infty$ and SECAM($\mathbf{X}$) $\leftarrow \varnothing$
for $t = 1$ to T **do**
  $\mathbf{p}_t \leftarrow \text{softmax}(M_t(\mathbf{X}))$
  $\mathbf{y}_t \leftarrow \arg\max_c \mathbf{p}_t[c]$
  if $\mathbf{y}_t = \mathbf{y}_{\text{ens}}$ **then**
    $\text{CAM}_t \leftarrow \text{ScoreCAM}(M_t, \mathbf{X}, \mathbf{y}_{\text{ens}})$
    $s_t \leftarrow \text{ADCC}(\mathbf{X}, \text{CAM}_t)$
    if $s_t > s^*$ **then**
      $s^* \leftarrow s_t$
      $\text{SECAM}(\mathbf{X}) \leftarrow \text{CAM}_t$
    end if
  end if
end for
return SECAM($\mathbf{X}$)

5 Experiments

Dataset: The Breast Ultrasound Dataset [1] used in this study contains breast ultrasound images from 600 female patients aged 25 to 75 years. The dataset was put together in 2018 and has 780 images, each with an average resolution of 500×500 pixels in PNG format. These images are divided into three classes: normal, benign, and malignant, with ground truth annotations provided for each image. This dataset provides a robust foundation for training and evaluating machine learning models for breast cancer classification, detection, and segmentation.

Model Architecture: The model architecture is based on the ResNet50 [3] network pre-trained on ImageNet. A custom classification head is added on top, consisting of three fully connected layers with 512, 128, and 3 neurons, respectively. A dropout layer with a probability of 0.2 is included between these layers to reduce overfitting. The Deep Ensemble model comprises 10 instances of this architecture, where each model is initialized with different random weights. During training, only the final convolutional layer of the ResNet50 backbone is fine-tuned, while the remaining layers are frozen.

Evaluation Protocol: As the primary focus of this study is on explainability within deep ensemble uncertainty quantification (UQ) methods, optimizing classification performance was not the main objective. Nevertheless, we evaluated

the classification performance of the model using standard metrics, including accuracy, precision, recall, and F1-score, to ensure a reasonably strong baseline. We then analyzed the relationship between classification performance and uncertainty estimates, specifically predictive entropy, in order to assess the ability of the UQ method to identify misclassified samples through elevated uncertainty. In addition, we evaluated the quality of the explainability maps generated by ECAM and SECAM by analyzing their alignment with ground truth masks, thereby assessing the reliability of the model's explanations. As a baseline, we considered the VCAM ensemble explainability method [11], which aggregates explanation maps from individual ensemble members using validation performance as weighting. This comparison allows us to qualitatively assess the effectiveness of the proposed methods.

5.1 Relationship Between Predictive Uncertainty and Model Performance

In our experiment, the Deep Ensemble model shows an accuracy of 88.61%, a precision of 89.46%, a recall of 88.61%, and an F1-score of 88.75%. These numbers suggest that the model's performance is good and similar to the best results on the Breast Ultrasound dataset. Additionally, we examine the relationship between predictive uncertainty and model performance (see Fig. 1).This helps us understand the practical value of predictive entropy derived from Deep Ensemble as a measure of uncertainty.

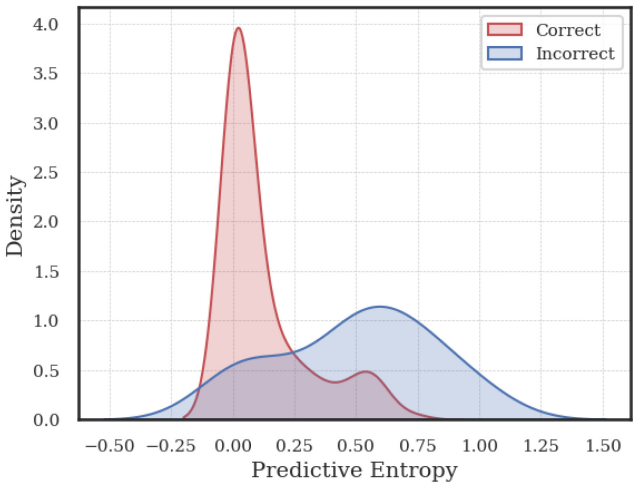


Fig. 1. Distribution of predictive entropy given classification performance.

A key observation is the difference between correctly classified and misclassified samples when examining predictive entropy. Correctly classified samples

have low predictive entropy, showing high confidence in predictions. In contrast, misclassified samples have high entropy, indicating greater uncertainty. This pattern shown through the density plots supports the value of predictive entropy as a measure of model confidence. These findings have several important implications. First, the link between high entropy and misclassified samples highlights the potential of predictive entropy as a tool for identifying unclear or difficult cases. In clinical settings, these cases could be identified for secondary review. This improves patient safety by reducing the dependence on uncertain predictions. This supports the broader goal of using risk-aware AI systems in sensitive areas like healthcare. Additionally, the results highlight opportunities for selective prediction. In this approach, the model avoids making predictions on samples with high uncertainty, referring them to human experts or other diagnostic tools.

Finally, misclassified samples with high uncertainty may reveal issues in the dataset, like edge cases or classes that are not well represented. Examining these situations further could result in specific improvements in both the dataset and the model training process.

5.2 Explanation Map Analysis

The qualitative analysis of the explainability maps presented in Fig. 2 and Fig. 3 provides detailed insight on how ensemble-based interpretability methods work in both correct and failure scenarios. In Fig. 2, where predictions are correct and the regions of interest (ROIs) are correctly identified, both ECAM and SECAM create activation maps that match the ground truth masks. When compared to the baseline VCAM, the proposed methods show more spatially focused and semantically consistent activations. This indicates that ECAM and SECAM better gather ensemble information to highlight clinically important areas, thereby improving how we interpret deep ensemble models.

Figure 3 on the other hand, shows two different types of failure modes. First, in the examples with red frames (correct classification but wrong localization), the model predicts the right label while concentrating on areas that don't match the annotated ROI. This behavior suggests that the model may depend on spurious or weakly correlated features that are predictive of the class but lack clinical significance. Alternatively, it may show that there are important visual cues that the ground truth annotations didn't pick up on. This indicates that the model can sometimes find complementary patterns that aren't in the masks that people defined. These situations emphasize the importance of explainability in understanding model behavior and improving UQ. Notably, ECAM usually generates more diffuse and less overconfident activation maps, similar to VCAM. In contrast, SECAM often produces sharper but possibly misleading explanations because it relies on just one ensemble member. This observation points out a trade-off between localization and predictive performance. It also promotes including uncertainty estimation alongside the model explanation. This helps assess how confident the model is in its interpretability results.

Second, in the green-framed examples (misclassification with correct localization), the explainability maps clearly show the pathological areas despite the

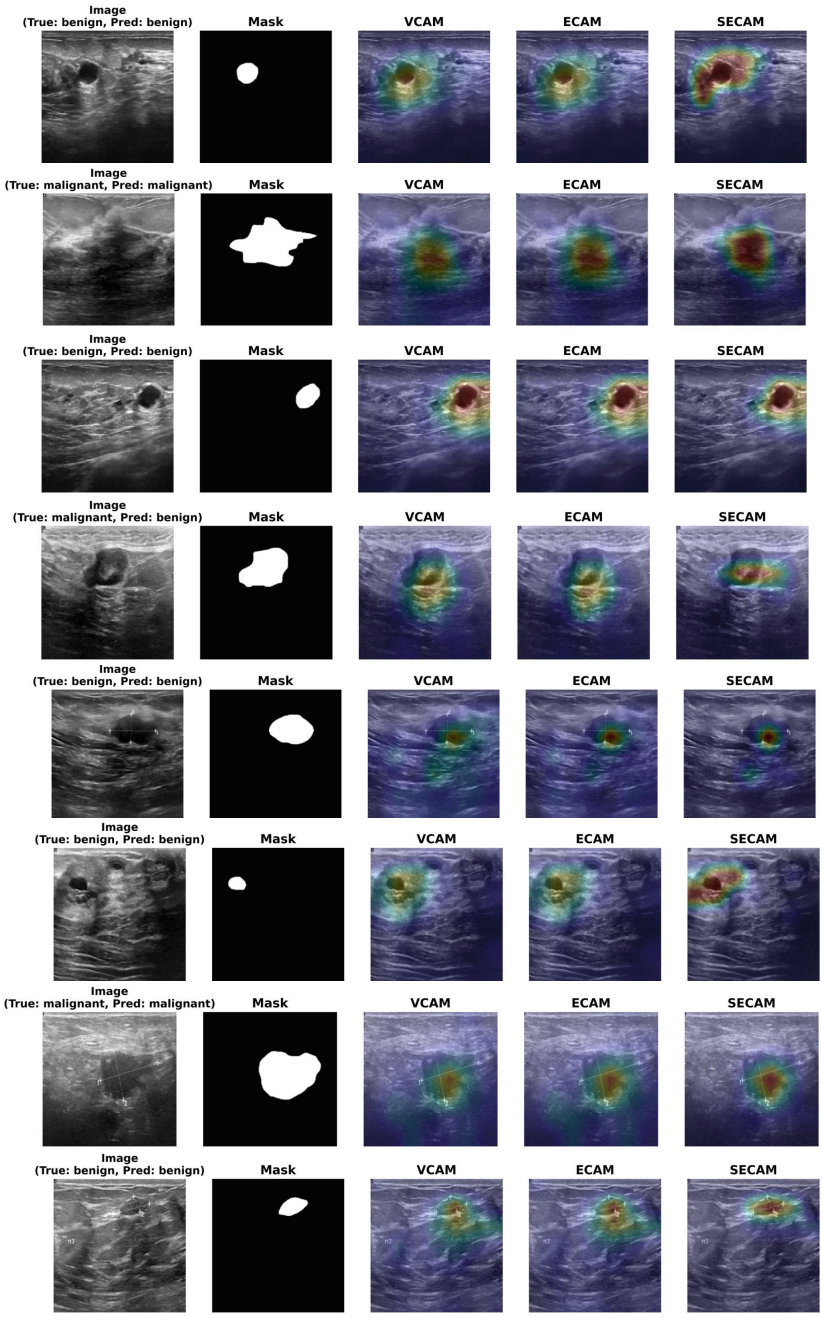


Fig. 2. Explainability map generated by ECAM and SECAM compared to VCAM on some correctly classified samples with accurate localization of regions of interest (ROIs).

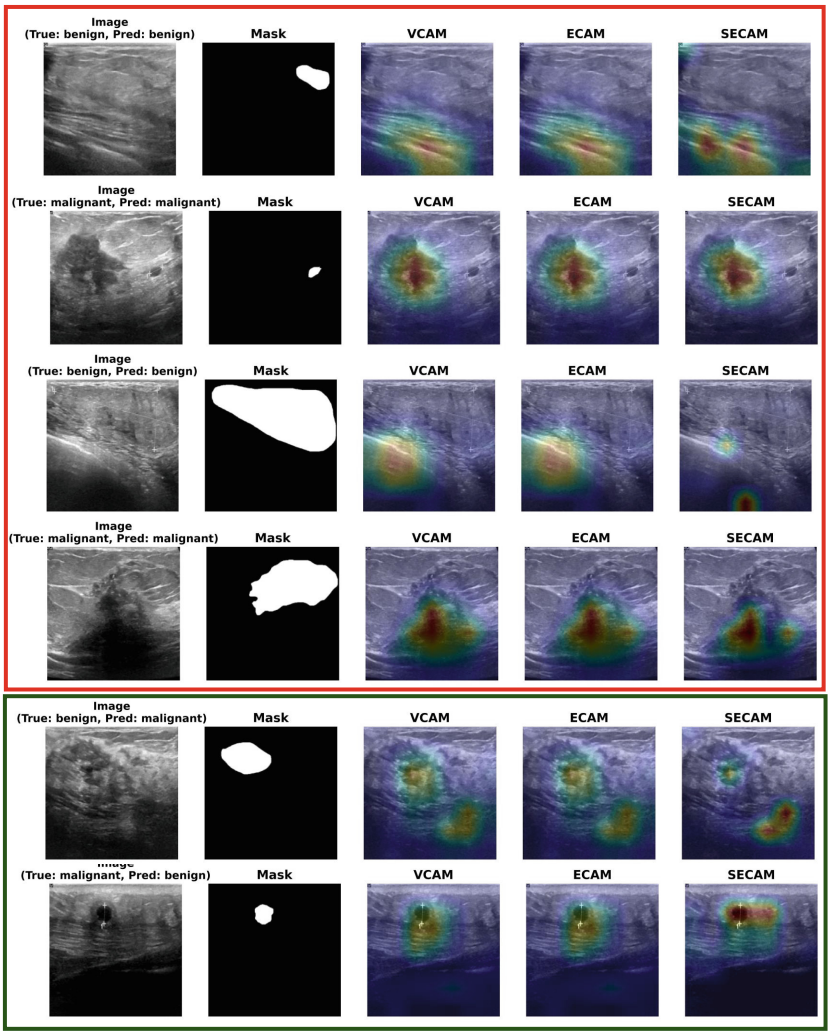


Fig. 3. Explainability failure modes. Red-framed examples show correct classifications with incorrect localization. Green-framed examples show misclassifications with correct localization. (Color figure online)

wrong predictions. This is an important observation because it suggests that the model can recognize key anatomical structures but struggles during the decision-making stage. This difficulty likely arises from confusion in feature representation or overlap among classes, like benign versus malignant. These cases show that classification errors do not always come from poor localization. Instead, they arise from limits in learning features that are discriminative. This difference is

important from a clinical point of view. Correct localization can still provide valuable diagnostic support, even if the predicted label is incorrect.

Overall, ECAM and SECAM show consistent and stable behavior, even though they use different methods to create ensemble-level explainability maps. The results show that adding explainability to deep ensemble frameworks makes the model more transparent and more informative. However, there is an important limitation: the quality of ensemble-level explanations depends on how reliable the individual ensemble members are. If the base models fail to provide reliable or meaningful explanations, the combined ensemble explanation cannot be completely trusted. This shows how important it is to improve model performance, estimate uncertainty, and enhance explainability. Doing so will help develop trustworthy AI systems for critical uses like medical imaging.

6 Limitations

Despite the promising results of this study, several limitations still exist. First, the algorithms were evaluated on a single dataset, and their generalizability to other imaging modalities and datasets has not yet been established. Second, the reliability of ensemble-level explanations is fundamentally dependent on the quality of individual ensemble members. This highlights a structural dependency that cannot be fully mitigated at the aggregation level alone and suggests that improving base model robustness is a prerequisite for reliable ensemble interpretability. Finally, the proposed methods rely on post-hoc explainability techniques, which inherit known limitations such as sensitivity to model architecture, gradient saturation, and lack of causal interpretability. As a result, the generated explanations should not be interpreted as causal evidence but rather as approximations of model attention, which may not always reflect true decision-making processes.

7 Conclusion

In this work, we introduced two novel ensemble-based explainability methods, ECAM and SECAM, designed to improve the explainability of Deep ensemble models used for uncertainty quantification in medical imaging. Our work was applied to the diagnosis of breast cancer using ultrasound images. By leveraging complementary aggregation strategies, the proposed methods provide insights into both collective ensemble behavior and individual model contributions, enabling a more comprehensive understanding of model decision processes. Through qualitative analysis, we demonstrated that ECAM and SECAM are capable of producing coherent and clinically relevant activation maps, while also revealing important failure modes such as the decoupling between localization and classification. Overall, this work contributes to the growing effort toward trustworthy AI by bridging ensemble learning, uncertainty quantification, and explainability. We believe that such integrated approaches are essential for enabling the safe and reliable deployment of deep learning systems in high-stakes domains such as medical imaging.

Future work will assess the generalizability of ECAM and SECAM across additional datasets and imaging modalities, including histopathology, cytology, mammography, and CT imaging. This assessment will help us evaluate their reliability in various clinical settings. We also plan to investigate uncertainty in explanations themselves, with the aim of further improving the reliability and transparency of model interpretation.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Gawlikowski, J., et al.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**(1), 1513–1589 (2023). https://doi.org/10.1007/s10462-023-10562-9
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
4. Hosny, K.M., Mohammed, M.A., Salama, R.A., Elshewey, A.M.: Explainable ensemble deep learning-based model for brain tumor detection and classification. Neural Comput. Appl. (2024). https://doi.org/10.1007/s00521-024-10401-0
5. Hryniewska-Guzik, W., Longo, L., Biecek, P.: CNN-based explanation ensembling for dataset, representation and explanations evaluation. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) Explainable Artificial Intelligence, vol. 2154, pp. 346–368. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-63797-1_18
6. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
7. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: a unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif. Intell. Med. **150**, 102830 (2024). https://doi.org/10.1016/j.artmed.2024.102830
8. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions (2017)
9. Ntrougkas, M., Gkalelis, N., Mezaris, V.: TAME: attention mechanism based feature fusion for generating explanation maps of convolutional neural networks. In: 2022 IEEE International Symposium on Multimedia (ISM), pp. 58–65 (2022). https://doi.org/10.1109/ISM55400.2022.00014
10. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting the evaluation of class activation mapping for explainability: a novel metric and experimental analysis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2299–2304 (2021). https://doi.org/10.1109/CVPRW53098.2021.00260
11. Rajani, N.F., Mooney, R.J.: Ensembling visual explanations. In: Escalante, H.J., et al. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 155–172. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_7
12. Ribeiro, M.T., 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). https://doi.org/10.1145/2939672.2939778

13. Rieger, L., Hansen, L.K.: Aggregating explanation methods for stable and robust explainability. arXiv: Learning (2019)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
15. Shorfuzzaman, M., Hossain, M.S., El Saddik, A.: An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading. ACM Trans. Multimedia Comput. Commun. Appl. **17**(3s) (2021). https://doi.org/10.1145/3469841
16. Sritharan, N., Gnanavel, N., Inparaj, P., Meedeniya, D., Yogarajah, P.: Ensemble-CAM: unified visualization for explainable cervical cancer identification. In: 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE), vol. 7, pp. 1–6 (2024). https://doi.org/10.1109/SCSE61872.2024.10550859
17. Stodt, J., Reich, C., Clarke, N.: A Novel metric for XAI evaluation incorporating pixel analysis and distance measurement. In: 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1–9. IEEE Computer Society (2023). https://doi.org/10.1109/ICTAI59109.2023.00009
18. Thakur, G.K., Thakur, A., Kulkarni, S., Khan, N., Khan, S.: Deep learning approaches for medical image analysis and diagnosis. Cureus **16**(5), e59507 (2024). https://doi.org/10.7759/cureus.59507
19. Vanitha, K., Mahesh T., R., Sree, S.S., Guluwadi, S.: Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning. BMC Med. Inform. Dec. Mak. **24**(1), 222 (2024). https://doi.org/10.1186/s12911-024-02628-7
20. Wang, H., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 111–119 (2020). https://doi.org/10.1109/CVPRW50498.2020.00020
21. Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Naseem, U., Gu, Y.: Robust and explainable framework to address data scarcity in diagnostic imaging (2024). https://doi.org/10.48550/ARXIV.2407.06566
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

A Laboratory Testbed for Benchmarking Orchestration Logic in Collaborative Environments

Leopoldo O. Silva[1,2(✉)], Diogo L. Rodrigues[3], Francisco J. Duarte[1,2], and Luís S. Dias[1,2]

[1] CCG/ZGDV Institute, University of Minho, Azurém Campus, 4800-058 Guimarães, Portugal
leopoldo.silva@ccg.pt
[2] ALGORITMI Research Centre, University of Minho, 4710–057 Braga, Portugal
[3] Colep Packaging, Rua Comendador Arlindo Soares de Pinho 1977, Vale de Cambra, Portugal
https://www.ccg.pt

Abstract. The need for agile and decentralized supply chains, demanded by Industry 4.0, requires seamless integration among several Enterprise Information Systems. On this paper we present an architecture and laboratory exploration tool designed to test and benchmark orchestration algorithms within a collaborative environment. By conceptualizing the Enterprise Information Systems as digital representations of the ecosystem enterprises and using a modern context broker for synchronization, our laboratory tool acts as a decoupled orchestration architecture. We validate this tool through a First-Come, First-Served baseline algorithm and quantify the architecture performance via a set of laboratory experiments measuring the orchestration latency, throughput, fidelity and architecture resilience. The results demonstrate that our architecture provide a low-latency, scalable and resilient foundation for coordinating supply chains experiments, confirming the viability of the proposed transactional shared data model and offering a pathway for further tests with advanced planning algorithms in diverse industrial contexts.

Keywords: Supply Chain Orchestration · Planning Algorithms · Industry 4.0 · Digital Twin

1 Introduction

Industry 4.0 requires agile, decentralized supply chains synchronized across heterogeneous Enterprise Information Systems (EIS) [1]. However, to achieve the needed operational agility, several challenges remain unsolved, being data silos and limited visibility between organizations two of those [2]. This research addresses these challenges by proposing a collaborative environment where each

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 177–192, 2027.
https://doi.org/10.1007/978-3-032-30494-0_12

participant is an enterprise, represented by its Digital Twin of the Organization (DTO). By leveraging the FIWARE Orion Context Broker and the NGSI-LD standard, a shared digital context is established by the framework, allowing orchestration among the participant enterprises. The main contribution is a laboratory tool designed to explore and benchmark orchestration within distinct industrial scenarios. This tool was validated through a First-Come, First-Served (FCFS) baseline within a fragmented enterprise ecosystem, providing a pathway for future tests on supply chain synchronization in distinct industrial scenarios.

2 Related Work

One the main goals of Industry 4.0, is to integrate Cyber-Physical Systems and the Internet of Things into manufacturing [3], with the Digital Twin serving as a real-time virtual mirror of physical assets [4]. The synchronization of decentralized value chains, with their own data silos, using a network of twins to achieve seamless collaboration, requires event-driven mechanisms capable of keeping low-latency responses during high-frequency orchestration tasks, a requirement that is not well quantified in decentralized industrial contexts [5].

Using EISs as the source of truth, DTOs abstract complete enterprises and their processes by keeping operational states up to date. [6]. However, even if these twins are already able to provide capacity and constraints models, there is still a research gap in how to federate this models for cross-company orchestration, particularly regarding mechanisms to persist shared digital states and maintain synchronization after failures. [7].

The FIWARE ecosystem [8] and the NGSI-LD standard [9] support the technical interoperability required by these federated environments. Context brokers manage high-density data flows by decoupling ecosystem participants through standardized Interfaces. The asynchronous notification architecture implemented by these technologies is essential for maintaining the fidelity of the DTOs and their digital shadow [10], ensuring that state changes, such as production progress, are propagated across the ecosystem without the delays associated with traditional methods.

To plan a decentralized supply chain, trade-offs between local optimization and global efficiency are required [11]. As such, a baseline logic, like the FCFS, provides a neutral foundation for benchmarking more advanced algorithms [12]. There is also a clear need for evidence-based testbeds that can empirically measure the throughput and reliability of these algorithms in diverse industrial scenarios.

3 Research Methodology

On this work we develop a laboratory testbed for benchmarking decentralized planning strategies, using the Design Science Research Methodology [13]. This methodology allows us to interactively design and evaluate artifacts in controlled environments, which is essential for bridging theoretical supply chain models

with the practical interoperability requirements of Industry 4.0. On the design phase, four critical requirements were identified: semantic interoperability to unify diverse EIS, state representation for digital twin synchronization, architectural scalability for concurrent requests, and system resilience to ensure state consistency during service initialization and restoration. The FIWARE ecosystem was selected as the technological stack for context-aware data management, since it promises to facilitate the resolution of those requirements. It also results from an European project and will allow the technological sovereignty that is sought off in current European directives.

A single-hub multi-spoke model was used on the experimental design to simulate coordination across decentralized participants while ensuring the isolation of variables. Õn this framework, several planning algorithms serve as independent variables, being our future goal to test distinct operational philosophies, ranging from basic sequencing to complex collaborative optimization [14]. At this step, the study uses a FCFS baseline to establish empirical benchmarks for system connectivity and asynchronous state machine transitions.

To validate our architecture we realized several simulations that cross-reference the orchestration logic with specific business pressures. A multi-metric evaluation matrix was used to quantify performance, focusing on orchestration latency, system throughput, twin convergence, and initialization recovery time. This systematic approach provided empirical evidence to assess the suitability of the orchestration logic for fragmented participant ecosystems. With the quantification of these parameters, the research demonstrates the utility of our foundational architecture for further experimentation on synchronized and resilient supply chain management.

4 The Experimental Simulation Framework

We have designed the experimental simulation framework as a modular and interoperable architecture, illustrated in Fig. 1, with the goal of facilitating the digital representation of complex supply chain dynamics. On it, each participant, represented by their EISs, can interact through a unified digital interface. By conceptualizing these systems as DTOs, the framework enables the mapping of internal operational states to a shared global context. This approach ensures that participant's information, such as production capacity, request status, and resource availability, is consistently maintained and synchronized across the ecosystem, providing a solid foundation for automated decision-making and strategic planning.

The integration of distributed data sources through a centralized context broker is the core of this framework and it serves as the synchronization backbone for all digital entities within the Digital Twin Layer. Using standardized communication protocols and semantic data models, the framework decouples the internal complexity of individual EISs, in the Enterprise layer, from the orchestration logic in the Service layer. This decoupling is essential for creating a flexible testbed where different planning paradigms can be evaluated against various business constraints.

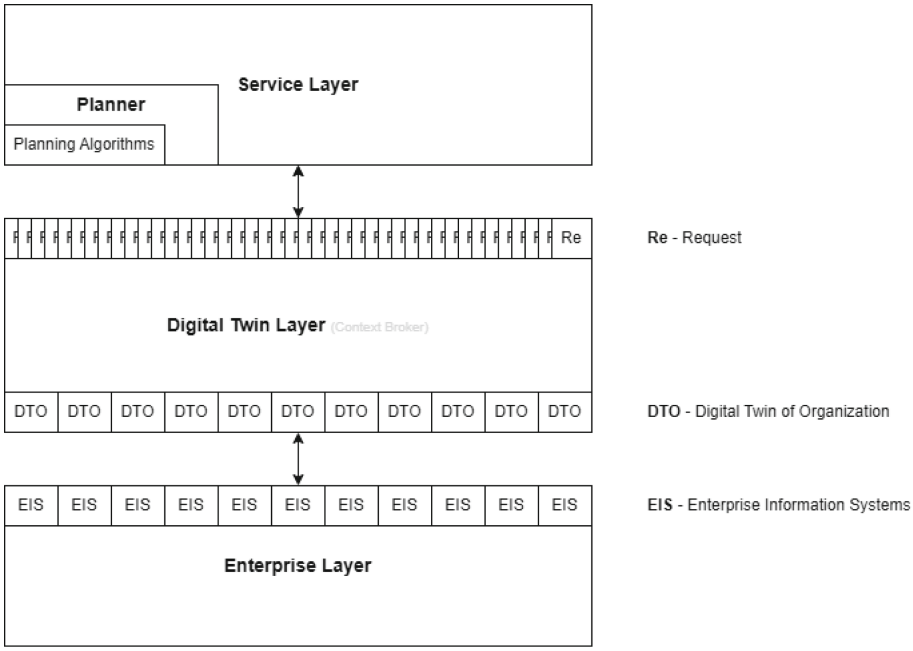


Fig. 1. High-level system architecture of the collaborative simulation framework.

The following sections detail the structural components of the framework, including the technological infrastructure, it's layers and the shared data model.

4.1 The Testbed Environment

To provide a context-aware and robust infrastructure for our industrial ecosystem, the FIWARE Orion context-broker was used as the foundation of the framework, function as a high-performance mediator of the testbed and maintaining a consistent global context across the digital twins. This approach promises to eliminate information gaps by ensuring that all participating entities share a synchronized view of the industrial state.

The broker maintains an up-to-date digital representations through the NGSI-LD standard, where participants and their requests, are modeled as semantic entities. This linked-data approach creates a comprehensive digital data set that serves as the single source of truth for the network, facilitating the complex inter-organizational coordination required for our experiments.

Furthermore, the architecture decouples data producers from consumers through an asynchronous publish and subscribe mechanism. The participants interact with each-other, solelely by the mediation of their digital twins and the broker, which automatically trigger notifications to the service layer when the right conditions are met. This event-driven interaction should allow high system

responsiveness and scalability, supporting the rapid state transitions that we expect to observe during experimental validation.

4.2 Data Models

To ensure seamless interoperability, a semantic approach based on the NGSI-LD standard is used. A common digital language is defined with the goal of describing the participants and their capacities. The data architecture centers on two primary linked entities: the Participant, representing the physical ecosystem with their EISs, and the Request, representing the lifecycle of the transactions among participants.

Table 1. Data Model Attributes for the *Participant* Entity

Attribute	Data Type	Description	Orchestration Utility
`id`	URI	Unique system identifier	Entity referencing
`name`	Property	Human-readable identifier	Identity and logging
`roles`	List	Consumer/Producer status	Capability filtering
`location`	GeoProperty	Geographic coordinates	Transport latency
`maxThroughput`	Property	Absolute capacity limit	Scalability assessment
`currentWorkload`	Property	Active request count	Fairness distribution
`availableCapacity`	Property	Remaining processing space	Concurrency control
`operationalStatus`	Property	Current availability state	Resilience/Re-planning
`avgProcessingTime`	Property	Manufacturing time per unit	Total Lead Time

The Participant entity is modeled as a multi-role object that supports industrial prosumer behaviors, by allowing participants to act as producers, consumers, or both. Attributes summarized in Table 1 provide the identity, spatial context, and operational states necessary for orchestration. The current experiments only use a subset of these fields, others such as `location` and `avgProcessingTime` are maintained to allow the extensibility of the platform for future logistical scenarios.

The Request was designed as a consolidated transactional artifact that integrates planning and execution data directly into the entity (Fig. 2). This model integrates the planning and execution phases into the request itself, defining sequential responsibilities for participants. The `status` attribute functions as a dynamic state machine, transitioning through stages such as `PENDING`, `ALLOCATED`, `IN_PRODUCTION`, and `COMPLETED` and allowing all involved participants to update their own internal states and with their own logic, by monitoring the evolution of the request (Table 2).

Our hope is that this merging of planning, execution, and lifecycle status into a single entity, allows the framework to ensure high data consistency and provide

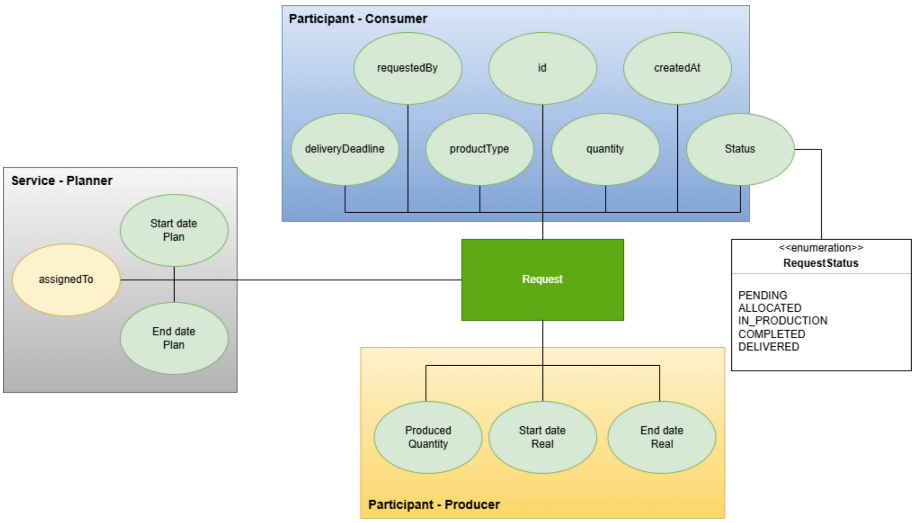


Fig. 2. Request entity data model: attributes, stakeholder responsibilities, and lifecycle status transitions.

Table 2. Data Model Attributes for the *Request* Lifecycle

Responsible	Attribute	Function
Consumer	`id, productType, quantity`	Demand Definition
	`deliveryDeadline, requestedBy`	Constraints and Origin
Planner	`assignedTo`	Allocation Decision
	`startDatePlan, endDatePlan`	Optimized Schedule
Producer	`producedQuantity`	Execution Monitoring
	`startDateReal, endDateReal`	Real-time Feedback
Shared	`status` (Enumeration)	Workflow Synchronization

an up-to-date view of the underlying processes. This integrated structure simplifies interoperability requirements for the participants while providing a scalable foundation for synchronized and resilient supply chain orchestration. Similarly to the participant entity model, the request model has specific attributes, such as planned and real execution dates, that are only partially used on the current baseline tests. These attributes were kept to enable future evidence-based decision-making through precise deviation analysis.

4.3 The Enterprise and Digital Twin Layer

The Digital Twin layer interfaces internal organizational logic with the collaborative network by representing each participant as a DTO. By utilizing the EIS as the foundational source of truth, this layer abstracts each participant business

complexities into standardized digital states. Functionally, this layer ensures that transactional changes in the EIS are reflected in the digital counterpart through an event-driven mechanism, allowing digital twins to act as mirrors of organizational constraints and facilitating the state convergence we wish to attain during experimental validation.

Finally, this layer supports information flows bi-directionally, acting both as a sensor and actuator for the participants EISs, either updating the status on the digital representation of the participant or sending orchestration commands to the EISs, This command injection into organizational workflows should allow the integration of the centralized assignments into local constraints.

4.4 The Service Layer

The service layers provides the specialized coordination processes of our testbed, representing the strategic command center of the environment. While the Orion Broker provides a dynamic mirror of the network's overall state, including request fulfillment and participant availability, this layer adds the centralized governance for the interactions between the participants of our framework.

At the moment and for the purpose of our experiments, the provided services are exclusively request fulfillment planners and allocators, allowing this layer to govern the industrial transaction lifecycle. Upon detecting a new request, the planner initiates the allocation process and triggers the necessary updates to the shared entities. Such governance ensures that interactions remain synchronized and aligned with the strategic objectives of the specific industrial scenario being simulated.

The service layer ensures system resilience without requiring direct integration with the EISs and this modular design allows for the seamless replacement of planning algorithms while maintaining a trail of all activities. By acting as the primary agent of change, the service layer transforms independent participants into a reactive and synchronized ecosystem.

5 Experimental Design and Setup

On this section we detail the laboratory-based evaluation we did to quantify the performance and reliability of the implemented framework. The goal was to validate the feasibility for synchronization and functional benchmarking by establishing a controlled simulation environment. Subsections 5.1 and 5.2 outline the FCFS orchestration baseline and the participant ecosystem parameterization, while Sect. 5.3 presents the experimental procedures used to measure orchestration latency, throughput, digital fidelity, and architectural resilience. With this approach we want to ensure that the results are based on empirical evidence and provide an assessment of the framework's suitability for decentralized supply chain experiments.

5.1 Implementation Baseline

Our functional baseline, the FCFS algorithm, uses event-driven sequential logic to process requests based on arrival time. This algorithm serves as a transparent gatekeeper for decentralized allocation, establishing essential connectivity between participants through a neutral and predictable task-release mechanism. It also provides a clean and neutral logic baseline, able to measure the performance limits of the NGSI-LD synchronization layer before introducing another algorithms in later experiments.

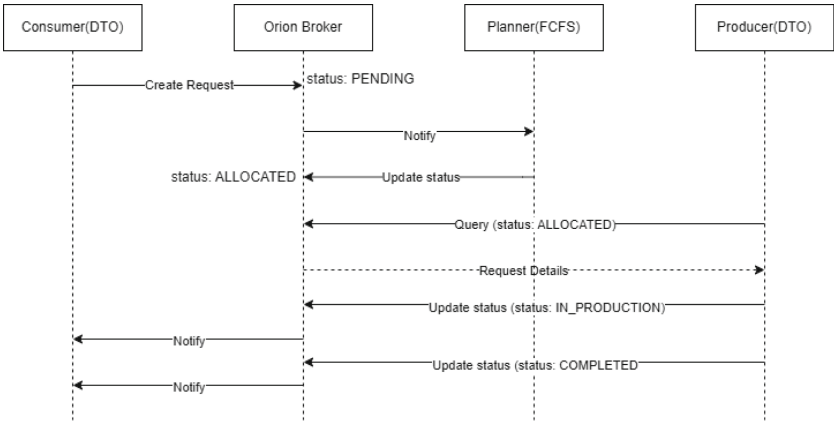


Fig. 3. Sequence diagram of the FCFS orchestration workflow.

Technical implementation relies on an asynchronous listener using the Orion Broker subscription system(Figure 3). The planner monitors requests in `PENDING` state and, upon notification, performs a state transition to `ALLOCATED`. This transition is the primary coordination signal, allowing producers to identify and claim tasks through their own decoupled notifications.

This shared context doesn't require point-to-point integration and enables a continuous feedback loop via automated status update notifications. As producers transition requests to `IN_PRODUCTION` and `COMPLETED`, notifications synchronize the consumers that created those requests. This asynchronous interaction among participants maintains a consistent state representation across all of them.

Furthermore, during the planner initialization, a synchronization routine executes to identify backlog requests created during orchestrator downtime. This query ensures that no transaction is overlooked, maintaining data consistency and a predictable state machine regardless of component startup timing or blackouts.

5.2 Simulation Environment

The simulation environment replicates a fragmented SME ecosystem, mirroring sectors where production is distributed across specialized low capacity par-

ticipants. Each SME is modeled as a DTO within the Orion Broker, featuring restricted processing windows and specific throughput constraints(resource modeling is parameterized by setting `maxThroughput` and `availableCapacity` to low threshold values). With this experiment we want to challenge the service layer to synchronize multiple small-scale units to satisfy collective demand.

5.3 Experimental Procedures and Metrics

Four laboratory experiments where done to evaluate the performance of the FCFS baseline within the SME ecosystem. The following sub-subsections detail protocols for assessing orchestration latency, system throughput, digital fidelity, and architectural resilience, providing a systematic framework to quantify the efficiency of our architecture for supply chain coordination.

Experiment 1: Orchestration Latency and Responsiveness. On this experiment we establish a baseline for orchestration latency, defined as the difference between initial request in a `PENDING` state and the following `ALLOCATED` update by the planner and received by the consumer. This metric assesses the responsiveness of the asynchronous notification mechanism, quantifying the architectural overhead. On this experiment there are only two participants, a consumer making requests and a producer receiving them.

The procedure involves creating 100 requests from a consumer, recording timestamps for broker processing, network latency, and logic execution. The mean latency and it's standard deviation are then calculated, allowing to evaluate the viability of the NGSI-LD standard for coordination. A low and consistent response time demonstrates that the decoupled architecture keeps agility and that the ecosystem can remain synchronized under strict temporal requirements.

Experiment 2: System Throughput Under Network Load. On this experiment we evaluate the scalability of the orchestration layer under high-density transactional flows. The goal is to assess the capacity of the Orion Broker and the FCFS planner to process bursts of information, ensuring that requests transition from `PENDING` to `ALLOCATED` without compromising context integrity.

The procedure tests the management of parallel threads and their database updates, simulating peak loads with a burst of 100 concurrent requests from multiple consumers. Using system throughput (measured in requests per minute) and the success rate of state transitions we identify potential bottlenecks or race conditions.

The framework's suitability for large-scale scenarios is determined by these limits. If the framework is able to achieve a high throughput rate, it may support the intense transactional demands of decentralized supply chains while maintaining operational efficiency.

Experiment 3: Digital Convergence and Fidelity. On this experiment we validate digital synchronization during the execution phase by focusing on the propagation of `producedQuantity` updates. The procedure simulates a production cycle where a producer advances the `producedQuantity` attribute, while the consumer updates is own state when notifications are received. The main metrics are synchronization jitter, measuring propagation time variance, and state fidelity, ensuring state consistency between participants (the consumer and the producer). By quantifying the convergence between local and perceived states, this experiment want to demonstrate the efficacy of the shared model for evidence-based decision-making.

Experiment 4: Resilience and Initialization Bootstrapping. On the final experiment we want to assess the resilience of the orchestration layer during component downtime, where resuming operations without data loss is crucial for reliability. The procedure tests the planner initialization routines by identifying requests missed during inactivity. For this, a participant injects a backlog of requests that remain in a `PENDING` state, during a controlled interruption of the planner. Upon restart, the initialization routine should obtain those requests and transition them to `ALLOCATED`. The main metrics are recovery time and chronological state consistency, which are important parameters for architectural robustness and the value of the broker's persistence layer during automated recovery in unpredictable environments.

6 Results

With the experimental procedures defined, as described in the previous section, we conducted the experiments and evaluation of our framework. The data collected with these experiments provided the empirical evidence regarding the performnace of the framework and the FCFS orchestration baseline within the fragmented SME ecosystem, On the following sections we present the quantitative results and technical observations obtained from each experiment, focusing on the efficiency of the asynchronous state transitions and the fidelity of the digital representations present on both the broker and the EIS simulations.

6.1 Experiment 1: Orchestration Latency and Responsiveness

The results for Experiment 1 quantify the end-to-end latency of the orchestration process, that is the elapsed time from the initial request creation by the consumer to the receipt of the request `ALLOCATED` status notification. This metric allows for the validating the reactivity of our framework. As shown in the representative sample (Table 3), the system demonstrated high temporal stability throughout the execution phase.

The system achieved an average end-to-end latency of 10.93 ms and the low standard deviation of 0.99 ms suggests that the asynchronous notification mechanism provided by the Orion Context-Broker, is highly predictable for coordination, ensuring that the overhead introduced by the digital twin and the service

Table 3. End-to-End Orchestration Latency Samples (ms)

Request ID	Latency (ms)
Request 01	9.98
Request 02	11.95
Request 03	10.13
Request 04	10.93
Request 05	10.28
Request 06	12.31
Mean Latency	**10.93**
Standard Deviation	**0.99**

layer remains well within the requirements for supply chain synchronization scenarios.

The logs also reveal that the primary component of this latency is on the internal dispatching logic of the context broker and the network round-trip between the distributed participants and services. The internal processing time of the planner service was consistently recorded as being below 2 ms, confirming that the FCFS logic is not a bottleneck. The consistency of these results validates the feasibility of the architecture for experimental scenarios and demonstrates that the decoupling of participants through a shared digital context maintains a high level of responsiveness.

6.2 Experiment 2: System Throughput Under Supply Chain Load

The objective of Experiment 2 was to evaluate the scalability and performance of the framework when subjected to a high volume of concurrent requests. The laboratory environment was configured with five consumers and ten producers, simulating a network of participants. The results (Table 4) show the system's capacity to stay under operational stability during peak load conditions.

Table 4. System Throughput and Load Performance Summary

Metric	Observed Value
Total Processed Requests	100
Simulation Duration (s)	80
Average Throughput (Requests/min)	75.0
Peak Throughput (Requests/s)	1.25
State Transition Success Rate	100%

These results show that the framework was able to hndle simultaneous injection of requests from multiple sources. As the consumers create requests, the

service layer processed the resulting notifications in a consistent way, ensuring that the shared context remained synchronized. A critical observation is the high degree of parallelism achieved at the execution level even if the FCFS planner acts sequentially, the decentralized participant layer enabled up to ten simultaneous producing cycles, effectively demonstrating some of the advantages of the decoupled architecture.

On the top of that, there is no evidence of race conditions or database deadlocks on the results. The stability of the throughput, which remained consistent at approximately 75 requests per minute, along with the previous findings, suggest that our architecture is capable of supporting the transactional density required for other decentralized scenarios and provides a robust foundation for more complex optimization algorithms in future research iterations.

6.3 Experiment 3: Digital Convergence and Fidelity

The objective of Experiment 3 was to validate the accuracy and synchronization of the digital representation during the execution phase. In our framework, fidelity is measured by the system's ability to propagate, with minimal delay, state changes from the producer to the consumer. By cross-referencing the timestamps of the updates done, by the producers, to the requests with the notifications at the consumers, we quantify the temporal gap in the digital representation of the shared context.

The results (Table 5), derived from the simulation involving 15 participants, indicate a high degree of temporal convergence, with a mean state synchronization latency of 8.42 ms. This suggests that the use of the Orion Context Broker for shared state management eliminates the information lag traditionally associated with inter-organizational communication. Even under the persistent load of many simultaneous production cycles, the jitter remained low and, as a result, the consumer's view of the producer's progress (`producedQuantity`) remained consistent and reliable.

Table 5. Digital State Convergence Latency (ms)

Event Type	Min Latency	Max Latency	Mean	Std. Dev.
Status: `STARTED`	4.12	14.56	9.15	1.34
Status: `PROGRESS`	3.85	12.10	7.82	0.95
Status: `COMPLETED`	4.30	13.92	8.29	1.12
Overall Average	–	–	**8.42**	**1.14**

These findings demonstrate the efficacy of the shared transactional data model, as the EISs of each organization may be synchronized across the network without the need for direct point-to-point integration. This may enable stakeholders to make evidence-based decisions, such as adjusting logistics or

downstream production schedules, based on the high-fidelity digital representation of their suppliers.

6.4 Experiment 4: Resilience and Initialization Bootstrapping

The final experiment evaluated the resilience of the architecture by assessing its ability to recover and synchronize pending requests after a period of downtime. This scenario is critical for environments where component failures cannot result in data loss or permanent supply chain desynchronization.

Table 6. System Resilience and Initialization Recovery Metrics

Metric	Observed Value
Backlog Size (Requests)	50
Total Recovery Time (ms)	421.0
Bootstrap Throughput (Req/s)	118.7
Transition to Live State	Seamless
Data Consistency Post-Recovery	100%

The results (Table 6) demonstrate that the initialization routine was able identify all requests created during the FCFS service inactivity. The framework shown a rapid recovery ability, having cleared the backlog with a high-performance burst and achieving a throughput of over 118 requests per second. This allowed the DTOs and the corresponding EISs simulations to be brought back into alignment with the global state almost instantaneously.

The experiment also shown a seamless transition between the initialization sweep and the asynchronous notification mechanism. Requests created just after the planner's restart were processed with the baseline latency from Experiment 1. These findings validate the robustness of our architecture, demonstrating that the use of a persistent context broker as a single source of truth provides the necessary resilience for experimental industrial deployments.

7 Discussion and Analysis

The empirical evaluation validates that our framework is ready for future experiments in decentralized supply chain orchestration. The following analysis interprets this results, specifically assessing the architectural trade-offs and strategic value of the collaborative environment within Industry 4.0 scenarios.

First, the conceptualization of EIS as DTOs provides a semantic context that unifies distinct data sources usually residing in their own data silos. By decoupling the internal complexity of the EISs from the orchestration logic, the framework avoids the cost of point-to-point integration while enabling participants synchronization. This transition changes the industrial ecosystem into a

unified system of systems, where participant states are available for coordinated decision-making.

FCFS results indicate that the context broker overhead is irrelevant when compared to industrial lead times. The observed 10 ms end-to-end latency proves the responsiveness of the architecture for real-time orchestration. Paired with the throughput of 75 requests per minute, it also validates system scalability for maintaining agility in fragmented networks.

Another key finding is the temporal convergence between physical execution and digital representation. The experiments on fidelity confirm a consistent digital representation across the participants, with synchronization jitter remaining in low ranges. This may enable evidence-based decision-making by eliminating operational blind spots and shows that shared transactional artifacts provide a reliable source of truth for producing life-cycles.

The strategic value of shared states for system resilience is shown by the bootstrapping results. With the achievement of a recovery throughput of 118 requests per second, the framework demonstrates the ability to handle backlog surges while preserving chronological integrity. This robust transition between recovery and notification-driven operations ensures that the framework survive the intermittent failures typical of distributed industrial environments.

Finally, even if FCFS establishes essential connectivity, its sequential logic lacks the logistical awareness required for complex supply chains. Our architecture addresses this by acting as a testbed for future advanced algorithms experiments, including Concurrent and Collaborative Planning. In this study, FCFS acts as a controlled baseline that allows us to isolate the architectural overhead from algorithmic complexity, ultimately validating that the framework is ready to host more advanced algorithms. We do believe that the underlying data models enable these kinds of algorithms and their application to other sectors and applications, where synchronized digital states are paramount for long-term operational sustainability.

8 Conclusion and Future Work

This research validates our experimental laboratory testbed designed to experiment with fragmented supply chains Industry 4.0 scenarios. By assuming that many of the current interconnection issues with EISs may be solved by conceptualizing them as DTOs and using FIWARE Orion context broker as a backbone to keep those DTOs data, our framework seems to enable real-time orchestration through a shared digital context. Our experimental validation using the FCFS baseline confirmed that the architecture as the needed responsiveness, scalability and resilience, proving that our common data model and respective instantiated entities maintain the needed fidelity for synchronizing independent organizations EISs. This laboratory testbed serves as a functional artifact for evidence-based decision-making, by providing a single source of truth via shared transactional artifacts and offering an agile, interoperable alternative to brittle point-to-point integrations.

Preliminary integration of logistically aware heuristics is already showing promising results, further demonstrating the modularity of the Service Layer for multi-criteria benchmarking. In future research we plan to experiment with advanced planning algorithms, such as Service Fairness for equitable load balancing, Concurrent Planning for high-density environments, and Collaborative Planning for multi-criteria matchmaking. These modules will leverage the transport and scheduling attributes, already in place in the data model, to address coordination challenges beyond sequential task release. Furthermore, the framework will expand to encompass new ecosystem scenarios like verticalized industries or global supply chains. Such extensions intend to demonstrate the versatility of our framework, positioning it as a foundational technology for resilient, synchronized, and sustainable global value chain experimentation.

Acknowledgments. This work has been supported by the European Union under the Next Generation EU, through a grant of the Portuguese Republic's Recovery and Resilience Plan (RRP) Partnership Agreement, within the scope of the project PRODUTECH R3 – "Agenda Mobilizadora da Fileira das Tecnologias de Produção para a Reindustrialização", Total project investment: 166.988.013,71 Euros; Total Grant: 97.111.730,27 Euros.

References

1. Hofmann, E., Rüsch, M.: Industry 4.0 and the current status as well as future prospects on logistics. Comput. Ind. **89**, 23–34 (2017)
2. Muñoz, M., Torres, M., Gil, J., Guzmán, J.: An internet of things platform for heterogeneous data integration: methodology and application examples. J. Netw. Comput. Appl. **240**, 104197 (2025)
3. Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., Hoffmann, M.: Industry 4.0. Bus. Inf. Syst. Eng. **6**, 239–242 (2014)
4. Tao, F., Zhang, H., Liu, A., Nee, A.Y.C.: Digital twin in industry: state-of-the-art. IEEE Trans. Industr. Inf. **15**(4), 2405–2415 (2019)
5. Fuller, A., Fan, Z., Day, C., Barlow, C.: Digital twin: enabling technologies, challenges and open research. IEEE Access **8**, 108952–108971 (2020)
6. Riss, U.V., Maus, H., Javaid, S., Jilek, C.: Digital twins of an organization for enterprise modeling. In: Grabis, J., Bork, D. (eds.) PoEM 2020. LNBIP, vol. 400, pp. 25–40. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63479-7_3
7. Kierans, D., Pleiter, D.: Realising distributed digital twins within federated digital infrastructures. In: DiDit 2024 - Proceedings of the 1st International Workshop on Distributed Digital Twins, co-located with DisCoTec 2024 - 19th International Federated Conference on Distributed Computing Techniques, DisCoTec 2024: CEUR-WS, 2024. QC 20250715 (2024)
8. FIWARE Foundation. Fiware ecosystem archives (2025). Accessed 30 Mar 2026
9. ETSI. ETSI GS CIM 009 V1.9.1: Context Information Management (CIM); NGSI-LD API (2025). Accessed 30 Mar 2026
10. Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W.: Digital twin in manufacturing: a categorical literature review and classification. IFAC-PapersOnLine **51**(11), 1016–1022 (2018). 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018

11. Saharidis, G.K., Dallery, Y., Karaesmen, F.: Centralized versus decentralized production planning. RAIRO - Oper. Res. **40**, 113–128 (2006)
12. Hesselink, W.H., Buhr, P.A., Parsons, C.A.: First-come-first-served as a separate principle. ACM Trans. Parallel Comput. **11** (2024)
13. Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S.: A design science research methodology for information systems research. J. Manag. Inform. Syst. **24**, 45–77 (2007)
14. An, Y., Zhang, Y., Zeng, B.: The reliable hub-and-spoke design problem: models and algorithms. Transp. Res. Part B Methodol. **77**, 103–122 (2015)

Metaheuristic-Optimized Ensemble Learning for Glioma Grading

Lucas Silva Santana, Saulo Moraes Villela(✉), Luiz Maurílio da Silva Maciel, Raul Fonseca Neto, and Marcelo Bernardes Vieira

Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, Brazil
lucas.santana@estudante.ufjf.br, {saulo.moraes,luiz.maciel,raulfonseca.neto,marcelo.bernardes}@ufjf.br

Abstract. Accurate grading of gliomas is essential for clinical decision-making, prognosis estimation, and treatment planning. Recent advances in molecular profiling have improved diagnostic accuracy, but their routine use remains limited due to cost, availability, and annotation constraints in public datasets. Consequently, machine learning approaches based on clinical and mutation features have gained increasing attention. However, no single classifier consistently achieves optimal performance across heterogeneous biomedical datasets. In this work, we propose a metaheuristic-optimized ensemble learning framework for glioma grading that integrates multiple classical machine learning models using weighted soft voting. Particle Swarm Optimization is employed to automatically determine the optimal contribution of each classifier in the ensemble, targeting improvements in accuracy, recall-oriented F2-score, and area under the ROC curve. In particular, emphasizing recall-oriented metrics is crucial in clinical settings, as false negatives, misclassifying high-grade gliomas as low-grade, may lead to delayed diagnosis and inadequate treatment, potentially resulting in worse patient outcomes. The framework is evaluated on two public datasets derived from The Cancer Genome Atlas and the Chinese Glioma Genome Atlas using a rigorous nested cross-validation protocol. On TCGA, the proposed method achieves an accuracy of 87.8% and an F2-score of 90.1%. On the CGGA dataset, the proposed approach attains an F2-score of 92.5% and an accuracy of 76.2%. Overall, the findings indicate that metaheuristic-driven ensemble optimization provides a flexible strategy for glioma grading, allowing the model to be adapted to different evaluation objectives, particularly in clinical scenarios where sensitivity to high-grade tumors is essential.

Keywords: glioma grading · brain tumor · clinical and molecular features · ensemble learning · metaheuristic optimization

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 193–208, 2027.
https://doi.org/10.1007/978-3-032-30494-0_13

1 Introduction

Gliomas are the most common primary tumors of the central nervous system and represent a major challenge in neuro-oncology due to their biological heterogeneity and variable clinical behavior [3]. Accurate tumor classification is a critical step in clinical decision-making, directly influencing treatment planning, therapeutic response assessment, and estimation of prognosis [16]. Gliomas are broadly categorized into low-grade and high-grade tumors, with glioblastoma multiforme (GBM) being the most aggressive form. Despite advances in diagnostic protocols, reliable classification remains a complex task due to overlapping histopathological features and variability between observers [1–3].

Recent updates in the World Health Organization (WHO) classification have incorporated molecular alterations, such as isocitrate dehydrogenase (IDH) mutation status, to improve diagnostic accuracy. Although these molecular markers provide significant clinical value, their routine use is constrained by high costs, long turnaround times, and limited availability of comprehensive clinical annotation in publicly accessible datasets. These challenges motivate the development of computational methods capable of leveraging existing clinical and molecular data more effectively while minimizing the use of extensive molecular testing [10,12].

Machine learning (ML) has demonstrated strong potential in medical decision-support systems by enabling data-driven, objective, and reproducible analyses. Classical ML models, including Support Vector Machines (SVM), Random Forests (RF), k-Nearest Neighbors (KNN), and gradient boosting-based methods such as Light Gradient Boosting Machine (LightGBM), have been widely applied to biomedical classification tasks due to their robustness and effectiveness in handling high-dimensional tabular data [9]. However, no single ML algorithm consistently achieves optimal performance across all datasets and problem settings, as formalized by the no-free-lunch theorem [15].

To address this limitation, ensemble learning approaches have been proposed to combine multiple classifiers, taking advantage of their complementary strengths to improve predictive performance and generalization [5,13]. Although conventional ensemble strategies such as bagging, boosting, and voting have achieved notable success, their effectiveness often depends on manually defined configurations and fixed combination schemes. Consequently, there is growing interest in adaptive and optimization-driven frameworks that can automatically determine optimal model combinations and configurations.

Metaheuristic optimization algorithms provide a powerful mechanism for navigating complex and high-dimensional search spaces and have been successfully applied to feature selection, hyperparameter tuning, and model combination tasks [6,8]. By integrating metaheuristics with classical ML models, it becomes possible to construct flexible ensemble systems that dynamically optimize classifier selection and combination strategies, leading to improved performance and robustness in heterogeneous datasets.

In this work, we propose a metaheuristic-based ensemble framework that integrates multiple classical machine learning models for glioma grading using

clinical and molecular characteristics. We employ Particle Swarm Optimization (PSO) [4] to optimize the weighting of individual classifiers within the ensemble, enabling objective-driven optimization tailored to specific evaluation criteria. In particular, the proposed framework allows prioritizing recall-oriented metrics, such as the F2-score, which are especially relevant in healthcare and clinical decision-making scenarios where false negatives may lead to delayed or inadequate treatment. In addition to recall-focused optimization, the framework can also target Accuracy and Area Under the Curve (AUC), providing flexibility to adapt the ensemble behavior to different clinical and methodological requirements. The PSO implementation used in this study is based on the Mealpy optimization library [14].

2 Related Work

Machine learning (ML) approaches for glioma grading and classification have been studied due to their potential to support clinical decision-making. Recent studies have explored a wide range of classifiers, feature selection strategies, and ensemble learning paradigms to address the heterogeneity and complexity of glioma grading.

Tasci et al. [12] proposed a hierarchical voting-based framework for glioma grading that integrates clinical and molecular features from The Cancer Genome Atlas (TCGA) and the Chinese Glioma Genome Atlas (CGGA) datasets. Their methodology employs multiple feature selection techniques, including Weight of Evidence, Recursive Feature Elimination, Random Forest-based importance, and Least Absolute Shrinkage and Selection Operator (LASSO), whose outputs are combined through a hierarchical voting scheme to identify robust and discriminative feature subsets. The selected features are subsequently evaluated using several supervised classifiers, namely Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Random Forest, and AdaBoost. Experimental results demonstrated that the proposed hierarchical voting-based feature selection strategy yields improved performance compared to the use of LASSO alone. They achieved an accuracy of 87.606% in the TCGA dataset and 79.668% in the CGGA dataset.

In a subsequent work, Tasci et al. [10] proposed the GradWise framework, which focuses on improving glioma grading through a hybrid feature selection strategy rather than ensemble classification. Their approach combines an embedded method (LASSO) with a filter-based method (mRMR) in a rank-based weighting scheme to identify the most relevant clinical and molecular features. They evaluated the selected feature subsets using five supervised models, including Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Random Forest, and AdaBoost. The experimental results showed that the proposed method achieved an accuracy of 87.007% in the TCGA dataset and 80.412% in the CGGA dataset.

Noviandy et al. [7] introduced a Light Gradient Boosting Machine (LightGBM) model integrated with Explainable Artificial Intelligence (XAI) techniques, specifically SHapley Additive exPlanations (SHAP), for glioma grading.

Their approach achieved a classification accuracy of 89.88% in the TCGA dataset by identifying key molecular features that strongly influence model predictions, thereby improving interpretability. However, their experiments were limited to the TCGA dataset, and no evaluation was conducted on the CGGA dataset, which restricted the assessment of the generalizability of the model among different populations. Furthermore, although the authors evaluated six different machine learning models for performance comparison, hyperparameter tuning was conducted exclusively for the LightGBM classifier, and they did not investigate ensemble learning or classifier combination strategies. It is also important to note that the experimental evaluation was performed using a single hold-out split, with 80% of the data used for training and 20% for testing, without employing k-fold cross-validation. As a result, the reported performance may be biased by the particular data partition used, making the results less robust and not directly comparable to approaches evaluated under cross-validation protocols, such as the method proposed in the present work.

As shown in Table 1, existing studies rely on fixed configurations or focus on feature selection without optimizing model combination or prioritizing clinically relevant metrics such as recall. To address these limitations, this work proposes a PSO-optimized ensemble that dynamically adjusts classifier contributions according to specific evaluation objectives, particularly in clinical scenarios where sensitivity to high-grade tumors is essential.

Table 1. Summary of related work.

Reference	Approach	Dataset(s)	Limitations
Tasci et al. [12]	Hierarchical voting-based ensemble combining multiple feature selection methods and classifiers	TCGA, CGGA	Lacks adaptive, objective-driven optimization mechanisms to adjust classifier contributions according to clinically relevant evaluation metrics, such as recall-oriented measures.
Tasci et al. [10]	Hybrid feature selection with rank-based weighting	TCGA, CGGA	Focuses on feature selection and does not explore ensemble learning optimization of multiple models
Noviandy et al. [7]	LightGBM model with explainability using SHAP	TCGA	Limited to a single dataset; lacks robust validation (no k-fold cross-validation); does not use ensemble methods or model combination optimization

3 Methodology

This section describes the data pre-processing steps, the ML models employed, the hyperparameter optimization strategy, the ensemble model combination approach based on metaheuristic optimization, and the evaluation metrics used in this study.

3.1 Datasets

In this work, we used two publicly available glioma datasets: the Glioma Grading Clinical and Mutation Features dataset derived from TCGA [11] and the dataset derived from CGGA [17].

The TCGA-based dataset was constructed to investigate the relationship between clinical characteristics, gene mutations, and glioma grading, specifically distinguishing between lower-grade gliomas (LGG) and glioblastoma multiforme (GBM). It consists of 839 samples described by 23 features, including 20 molecular and 3 clinical variables. The molecular features are binary, indicating the presence or absence of mutations in specific genes, and are encoded as 0 (not mutated) or 1 (mutated). Clinical features include gender, age, and race. Gender is encoded as 0 for male and 1 for female. Race is encoded as 0 for White, 1 for Black or African American, 2 for Asian, and 3 for American Indian or Alaska Native.

The molecular features of the TCGA dataset comprise the following genes: IDH1, TP53, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, IDH2, FAT4, and PDGFRA.

In addition, a second dataset derived from the CGGA was used. The CGGA dataset contains 286 samples described by 22 features. In contrast to the TCGA dataset, the CGGA dataset does not include the race feature, as all samples correspond to individuals from the Chinese population. Apart from this difference, the feature definitions and encoding schemes are consistent with those of the TCGA dataset.

For both datasets, glioblastoma multiforme (GBM) was defined as the positive class (label 1), while lower-grade glioma (LGG) was defined as the negative class (label 0) for all performance metrics evaluated.

3.2 Pre-processing

Prior to model training and evaluation, a pre-processing step was applied to ensure numerical stability and feature comparability. The numerical feature age was normalized using Min–Max normalization, scaling its values to the $[0, 1]$ interval. To prevent data leakage, the normalization parameters were computed exclusively on the training data within each outer cross-validation fold and subsequently applied to the corresponding test fold. This procedure is particularly important for distance-based and margin-based classifiers, such as k-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which are sensitive to

feature scaling. In the original dataset, the categorical variable race is encoded as an integer-valued feature. However, in our pre-processing pipeline, this variable is treated as categorical and transformed using one-hot encoding to avoid introducing artificial ordinal relationships between categories. All remaining features are binary and were therefore not subjected to additional scaling or transformation. In this study, we did not apply data balancing or resampling techniques.

3.3 Machine Learning Models

We conducted experiments using a diverse set of classical supervised machine learning models that are widely adopted in biomedical classification tasks. The evaluated models include Logistic Regression, Support Vector Machine with RBF kernel (SVM RBF), k-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost, Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Naive Bayes. This selection enables the evaluation of both linear and nonlinear decision boundaries, as well as tree-based ensemble methods, boosting techniques, and probabilistic classifiers, providing complementary perspectives on the underlying data distribution.

3.4 Hyperparameter Tuning

To identify the optimal hyperparameter configurations for each model, we employed a random search strategy. The random search was selected due to its efficiency in exploring high-dimensional hyperparameter spaces compared to exhaustive grid search. The hyperparameter search space for each model is detailed in Table 2.

We evaluated the dataset using a nested cross-validation strategy. The outer loop consisted of a 10-fold cross-validation used to estimate the generalization performance, where, at each iteration, the data were split into training and test subsets. Within each training subset, we performed an inner 5-fold cross-validation for model selection and hyperparameter tuning, using the F2-score as the optimization criterion. In this inner loop, the training data were further split into training and validation sets, and the validation set was used exclusively for hyperparameter optimization.

The optimized model configurations were then evaluated on the corresponding test fold. This nested evaluation strategy ensures that each sample in the dataset is used for testing exactly once, reducing dependency on a single train-test split and mitigating the risk of overfitting. Consequently, this approach provides a more reliable and unbiased estimate of the model's performance.

3.5 Model Combination Strategy

To further improve classification performance, we employed ensemble learning strategies to combine the predictions of multiple base classifiers. Initially, both soft voting and hard voting ensembles were evaluated. In soft voting, the predicted class probabilities of the individual classifiers are averaged with equal

Table 2. Hyperparameter search space explored for each model using random search, specifying value ranges, discretization steps, and sampling strategies (uniform and logarithmic scales).

Algorithm	Hyperparameter	Search space
AdaBoost	Number of estimators	[50, 500] (step = 50)
	Learning rate	[0.01, 1.0] (uniformly spaced, 5 values)
Decision Tree	Criterion	{gini, entropy}
	Maximum depth	[10, 30] (step = 10)
	Maximum features	{sqrt, log2}
KNN	Number of neighbors	[5, 25] (step = 5)
	Weights	{uniform, distance}
	Algorithm	{auto, brute}
LightGBM	Number of estimators	[50, 500] (step = 50)
	Minimum child samples	[3, 15] (step = 3)
Logistic Regression	Penalty	{l1, l2}
	C	[0.1, 100] (uniformly spaced, 25 values)
Naive Bayes	Var smoothing	$[10^{-12}, 10^{-3}]$
Random Forest	Number of estimators	[50, 500] (step = 50)
	Maximum features	{sqrt, log2}
SVM (RBF)	C	$[10^{-3}, 10^{3}]$ (log-scale, 10 values)
	Gamma	$[10^{-3}, 10^{1}]$ (log-scale, 5 values)
XGBoost	Number of estimators	[50, 500] (step = 50)
	Learning rate	[0.01, 1.0] (uniformly spaced, 5 values)

weights, allowing the final decision to incorporate model confidence. In contrast, hard voting determines the final prediction based on the majority of discrete class labels produced by the individual models.

Since individual classifiers may exhibit different levels of reliability, a weighted soft voting ensemble was also investigated. In this approach, a weight is assigned to each classifier, and the final prediction is obtained as a weighted combination of the predicted class probabilities, i.e., $\sum_{i=1}^{M} w_i \cdot p_i$, where p_i denotes the predicted probability of the positive class produced by the i-th classifier. The weights are constrained such that $\sum_{i=1}^{M} w_i = 1$. Algorithm 1 describes the training and model combination process.

The optimal set of weights was determined using the PSO metaheuristic, with the objective of maximizing specific performance metrics, namely Accuracy (ES ACC), F2 (ES F2) and AUC (ES AUC). PSO is a population-based optimization algorithm in which each particle represents a candidate solution that iteratively moves through the search space according to its own experience and that of neighboring particles.

Algorithm 1. Training and Evaluation with Nested Cross-Validation

Input: Dataset (X, y), set of base models $\mathcal{M}$
Output: Predictions for individual and ensemble models

```
Define 10-fold outer cross-validation
Define 5-fold inner cross-validation
for each outer fold do
    for each base model m ∈ M do
        Perform Randomized Search with inner CV optimizing F2-score
        Select the best model configuration
        Generate and store predictions on the outer test set
    end for
    Apply soft and hard voting ensembles and store outer test set predictions
    Optimize ensemble weights using PSO on the outer training data
    Apply optimized weights to the outer test set predictions and store results
end for
Compute evaluation metrics using aggregated predictions across all outer folds
```

At each iteration, the velocity of a particle is updated according to:

$$\mathbf{v}_i^{t+1} = w\mathbf{v}_i^t + c_1 r_1(\mathbf{p}_i - \mathbf{x}_i^t) + c_2 r_2(\mathbf{g} - \mathbf{x}_i^t), \tag{1}$$

where $\mathbf{v}_i^t$ and $\mathbf{x}_i^t$ denote the velocity and position of particle i at iteration t, respectively; $\mathbf{p}_i$ is the personal best position found by the particle; $\mathbf{g}$ is the global best position found by the swarm; w is the inertia weight; c_1 and c_2 are the cognitive and social coefficients; and r_1, r_2 are random numbers uniformly distributed in $[0, 1]$.

The particle position is then updated as:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1}. \tag{2}$$

The default parameter settings of the PSO implementation were used, namely epoch = 1000, population size = 50, $c_1 = 2.05$, $c_2 = 2.05$, and $w = 0.4$. These parameters control the balance between exploration and exploitation during the optimization process.

In our approach, each particle in the PSO population represents a candidate solution encoding the weights assigned to the base classifiers in the weighted ensemble. A particle is defined as a real-valued vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)$, where each component w_i corresponds to the weight associated with the i-th classifier, and M denotes the total number of base models. Therefore, the dimensionality of each particle is equal to the number of classifiers in the ensemble.

The optimization process was conducted exclusively on the training data within the nested cross-validation framework to avoid information leakage. Once the optimal weights were identified, they were fixed and applied to the independent test data, ensuring a fair and unbiased evaluation of the generalization performance of the weighted ensemble.

3.6 Evaluation Metrics

The performance of our model was assessed using multiple evaluation metrics to provide an analysis of its classification effectiveness. These metrics include accuracy, precision, recall, F1-score, F2-score, and area under the receiver operating characteristic curve (AUC). Although accuracy provides an overall measure of correctness, precision and recall capture complementary aspects of classification performance, particularly in the presence of class imbalance. The F1-score balances precision and recall equally, whereas the F2-score emphasizes recall, which is often critical in medical decision-making contexts. The AUC metric evaluates the model's discriminative ability across varying classification thresholds.

3.7 Experimental Setup

All experiments were conducted on a computer running Linux Mint 21.3, equipped with an Intel Core i7-12700 CPU (12th generation) operating at 2.10 GHz and 32 GB of RAM. We implemented all experiments in Python version 3.11, using scikit-learn version 1.7.2, LightGBM version 4.6.0, and the mealpy optimization library version 3.0.3. To ensure reproducibility, we fixed the random seed to 1 across all experiments.

4 Results

Initially, we conducted experiments using individual machine learning classifiers trained independently, without any ensemble or voting strategy. For each classifier, we performed hyperparameter optimization using the nested cross-validation approach described previously. Table 3 presents the results of the individual models evaluated on the TCGA dataset in the present study.

Table 3. Results of Individual Models on the TCGA dataset.

Model	ACC	PRE	REC	F1	F2	AUC
AdaBoost	86.65	79.27	**92.33**	85.30	**89.38**	90.20
Decision Tree	85.82	78.48	91.19	84.36	88.33	89.72
KNN	86.17	78.64	92.05	84.82	89.01	**90.99**
LightGBM	86.29	78.83	92.05	84.93	89.06	90.09
Logistic Regression	**87.60**	**82.12**	90.06	**85.91**	88.35	90.68
Naive Bayes	86.17	80.26	88.92	84.37	87.04	88.33
Random Forest	87.13	81.28	90.06	85.44	88.15	89.98
SVM RBF	86.65	79.41	92.05	85.26	89.21	90.19
XGBoost	85.46	78.33	90.34	83.91	87.65	90.25

All evaluated models achieved consistent performance, with accuracy values above 85% and AUC scores close to 0.90, indicating good overall discrimination capability.

In addition to accuracy and AUC, recall and F2 metrics were emphasized due to their relevance in clinical decision-making. AdaBoost, KNN, LightGBM, and SVM with the RBF kernel achieved recall values above 92%, which are reflected in their higher F2-scores. These results indicate a good ability to identify positive cases while prioritizing sensitivity. Although Logistic Regression and Random Forest achieved higher accuracy and precision, their recall and F2-score values remained competitive, indicating a balanced performance between precision and recall.

Table 4 presents the performance of the individual classifiers on the CGGA dataset.

Table 4. Results of Individual Models on the CGGA dataset.

Model	ACC	PRE	REC	F1	F2	AUC
AdaBoost	65.73	71.11	82.90	76.56	80.24	59.49
Decision Tree	65.73	69.71	87.05	77.42	82.92	65.24
KNN	71.68	75.93	84.97	80.20	83.00	71.52
LightGBM	71.33	74.89	86.53	80.29	83.92	71.19
Logistic Regression	70.63	72.80	**90.16**	**80.56**	**86.05**	65.17
Naive Bayes	**72.38**	**78.50**	81.35	79.90	80.76	**74.66**
Random Forest	68.88	77.08	76.68	76.88	76.76	71.07
SVM RBF	68.88	73.21	84.97	78.66	82.33	61.78
XGBoost	67.13	72.20	83.42	77.40	80.90	66.93

Overall, the results indicate moderate classification performance, with accuracy values ranging from 65.73% to 72.38% and AUC scores varying between 59.49 and 74.66, suggesting a more challenging classification scenario compared to the TCGA dataset.

With respect to recall-focused metrics, Logistic Regression achieved the highest recall (90.16%) and the best F1 and F2-scores, indicating a strong sensitivity to positive cases. Decision Tree and LightGBM also exhibited high recall values above 86%, which is reflected in their competitive F2-scores. Naive Bayes obtained the highest accuracy (72.38%), precision (78.50%), and AUC (74.66).

Overall, the results show that different models emphasize different performance aspects on the CGGA and TCGA datasets. These findings motivate the evaluation of ensemble strategies to combine complementary model behaviors and improve overall performance.

Table 5 presents the results of the ensemble strategies evaluated on the TCGA dataset, including soft voting, hard voting, and weighted ensemble models optimized using PSO. AUC was not computed for the hard voting scheme, as it produces only discrete class labels rather than continuous confidence scores required

for ROC analysis. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. We constructed the ensemble combinations using three different sets of base classifiers derived from Table 3: all evaluated models, a subset of five models (AdaBoost, KNN, LightGBM, Logistic Regression, and Naive Bayes), and a subset of four models (AdaBoost, KNN, Logistic Regression, and Naive Bayes).

Table 5. Ensemble Learning Results on the TCGA dataset.

Models	Method	ACC	PRE	REC	F1	F2	AUC
All	Soft Voting	87.01	80.60	90.91	85.45	88.64	91.31
	Hard Voting	87.01	80.60	90.91	85.45	88.64	–
	ES ACC	87.60	81.16	91.76	86.13	89.42	91.10
	ES F2	87.37	80.45	92.33	85.98	89.68	91.27
	ES AUC	87.13	80.20	92.05	85.71	89.40	91.28
5	Soft Voting	87.13	80.81	90.91	85.56	88.69	91.04
	Hard Voting	86.89	80.40	90.91	85.33	88.59	–
	ES ACC	**87.84**	**81.25**	92.33	**86.44**	89.88	91.30
	ES F2	86.89	79.95	91.76	85.45	89.13	90.82
	ES AUC	87.25	80.40	92.05	85.83	89.45	91.31
4	Soft Voting	87.13	80.81	90.91	85.56	88.69	91.09
	Hard Voting	86.77	80.66	90.06	85.10	88.01	–
	ES ACC	87.60	81.16	91.76	86.13	89.42	91.51
	ES F2	87.60	80.54	**92.90**	86.28	**90.13**	91.39
	ES AUC	87.60	80.69	92.61	86.24	89.96	**91.55**

The ensemble-based approaches exhibit performance comparable to, and in most cases better than, the individual classifiers.

Both hard and soft voting achieved comparable results in terms of accuracy, precision, recall, and F2-score, while soft voting achieved a higher AUC (91.31) than the best individual classifier (90.99%), indicating improved class separation. The PSO-based ensembles further enhanced performance by optimizing classifier weights according to different objective functions. The ensemble optimized for accuracy achieved the highest overall accuracy (87.84%) and F1-score (86.44%), while the ensemble optimized for F2 attained the highest recall (92.90%) and F2-score (89.68%), reinforcing its effectiveness in recall-sensitive scenarios. The ensemble optimized for AUC achieved the highest AUC among all evaluated combinations (91.55), reflecting improved class discrimination across decision thresholds. This performance surpasses the best individual classifier (AUC of 90.99), confirming the benefit of the ensemble approach for enhancing discriminative capability.

Overall, these results highlight the flexibility of the PSO-based ensemble framework, which allows for prioritizing different performance criteria while maintaining balanced and competitive classification performance.

Table 6 reports the performance of ensemble strategies on the CGGA dataset, including soft voting, hard voting, and PSO-based weighted ensembles optimized for accuracy, F2-score, and AUC. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. Compared to the TCGA dataset, the overall performance on the CGGA dataset is lower across all metrics, confirming the increased difficulty of this classification task.

Table 6. Ensemble Learning Results on the CGGA dataset.

Models	Method	ACC	PRE	REC	F1	F2	AUC
All	Soft Voting	73.43	74.89	91.19	82.24	87.39	73.55
	Hard Voting	71.68	74.56	88.08	80.76	85.00	–
	ES ACC	73.08	75.89	88.08	81.53	85.34	73.02
	ES F2	**76.22**	74.51	**98.45**	**84.82**	**92.50**	73.97
	ES AUC	72.03	75.34	87.05	80.77	84.42	69.30
5	Soft Voting	73.78	74.58	92.75	82.68	88.44	**75.67**
	Hard Voting	74.83	74.69	94.82	83.56	89.97	–
	ES ACC	72.38	76.64	84.97	80.59	83.16	72.98
	ES F2	71.33	71.60	95.34	81.78	89.41	72.88
	ES AUC	70.98	**77.50**	80.31	78.88	79.73	74.15
4	Soft Voting	74.13	73.71	95.85	83.33	90.42	74.87
	Hard Voting	73.78	75.00	91.71	82.52	87.80	–
	ES ACC	71.68	76.17	84.46	80.10	82.66	72.07
	ES F2	69.23	70.27	94.30	80.53	88.26	71.72
	ES AUC	71.33	76.81	82.38	79.50	81.21	74.07

Among the PSO-based ensembles, the model optimized for F2-score obtained the best overall results, achieving the highest accuracy (76.22%), recall (98.45%), F1-score (84.82%), and F2-score (92.50%), emphasizing its effectiveness in recall-oriented optimization. These results indicate that recall-oriented optimization effectively emphasizes the correct identification of high-grade gliomas, which is particularly important in clinical decision-making scenarios where false negatives may lead to delayed or inadequate treatment.

In contrast, the ensemble optimized for accuracy did not outperform the voting-based methods on CGGA. When compared to the TCGA results, the CGGA ensembles exhibit smaller absolute gains over voting strategies, indicating that the benefits of weighting classifiers are more dataset-dependent and more pronounced in the TCGA scenario.

Table 7 presents the results of the best-performing ensemble optimized for the F2-score on the TCGA dataset and compares it with state-of-the-art methods.

We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. The proposed approach employs PSO to determine optimal weights for each classifier, with a particular focus on improving the identification of positive cases when optimizing for the F2-score, while maintaining competitive overall performance. In addition, the table also reports the best results obtained by ensembles optimized for accuracy and AUC, illustrating the flexibility of the proposed framework in addressing different performance objectives.

Table 7. Comparison of our approach with state-of-the-art methods on the TCGA dataset.

Method	ACC	PRE	REC	F1	F2	AUC
Tasci et al. [12]	87.6	**85.8**	81.5	85.8	82.3	91.4
Tasci et al. [10]	87.0	80.4	91.5	85.5	89.0	91.1
ES ACC	**87.8**	81.3	92.3	**86.4**	89.9	91.3
ES F2	87.6	80.5	**92.9**	86.3	**90.1**	91.4
ES AUC	87.6	80.7	92.6	86.2	90.0	**91.6**

The main objective of our method is to prioritize the correct classification of the positive class while preserving competitive accuracy. When compared to the hierarchical approach proposed in Tasci et al. [12], the ensemble optimized for the F2-score achieves substantially higher recall and a higher F2-score for the positive class while maintaining equal accuracy. In comparison with the grad-wise method presented in Tasci et al. [10], the proposed ensemble optimized for the F2-score attains improved performance across all reported metrics. Moreover, the ensemble optimized for AUC outperforms both state-of-the-art methods, demonstrating superior overall discriminative capability. Finally, the ensemble optimized for accuracy achieves a higher accuracy than both reference methods, while also yielding higher recall, F1-score, and F2-score, further highlighting the effectiveness and flexibility of the proposed ensemble framework.

Overall, the higher recall, F2-score and AUC obtained by the proposed method demonstrate that the PSO-based weighting strategy effectively emphasizes the positive class, enabling improved sensitivity while maintaining competitive performance on the TCGA dataset.

Table 8 presents the results the best-performing ensembles on the CGGA dataset, optimized with respect to the F2-score, and compares it with state-of-the-art methods. We highlight the best results achieved in terms of accuracy, precision, recall, F1-score, F2-score, and AUC. The table also reports the best results obtained by ensembles optimized for accuracy and AUC. The experimental results indicate that the CGGA dataset represents a more challenging classification problem, as reflected by the lower accuracy.

Despite the lower accuracy, the proposed method demonstrates substantially superior performance in recall-oriented metrics, such as F2-score. Our method

Table 8. Comparison of our approach with state-of-the-art methods on the CGGA dataset.

Method	ACC	PRE	REC	F1	F2	AUC
Tasci et al. [12]	79.7	76.4	59.3	64.8	62.1	78.8
Tasci et al. [10]	**80.4**	**80.7**	60.7	67.9	63.9	**79.8**
ES ACC	73.1	75.9	88.1	81.5	85.3	73.0
ES F2	76.2	74.5	**98.5**	**84.8**	**92.5**	74.0
ES AUC	71.0	77.5	80.3	78.9	79.7	74.2

explicitly allows optimizing the ensemble toward different clinical priorities. In particular, the optimization of the F2-score emphasizes recall, which is highly relevant in clinical decision-making for glioma grading. In this context, false negatives (i.e., misclassifying high-grade gliomas as low-grade) may lead to delayed or inadequate treatment, potentially resulting in significantly worse patient outcomes. Therefore, improving recall from approximately 60% to up to 98.5% on CGGA represents a substantial clinical gain, even at the expense of moderate reductions in accuracy and AUC. These results highlight the effectiveness of the proposed PSO-based ensemble in emphasizing sensitivity to the positive class.

5 Conclusion

In this work, we proposed a metaheuristic-based ensemble learning framework for glioma grading using clinical and molecular features. By optimizing ensemble weights with Particle Swarm Optimization (PSO), the proposed method enables objective-driven optimization tailored to different evaluation criteria, including Accuracy, F2-score, and AUC.

Results on the TCGA and CGGA datasets showed that the proposed ensemble outperforms individual classifiers and conventional voting strategies, particularly when optimized for the F2-score, achieving higher recall and F2-score values while maintaining competitive accuracy. Although the CGGA dataset presents a more challenging classification task due to its reduced sample size, the proposed method preserved robust recall-oriented performance across both cohorts.

Overall, these findings suggest that metaheuristic-optimized ensembles offer a practical and adaptable solution for glioma grading using structured clinical and molecular features. Although evaluated on glioma grading, the proposed framework is applicable to other recall-sensitive classification tasks. Future work can explore larger and more feature-rich datasets to further evaluate the robustness and generalization of the proposed approach.

Acknowledgments. The authors would like to thank the Minas Gerais State Research Foundation - FAPEMIG.

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

References

1. Eder, K., Kalman, B.: Molecular heterogeneity of glioblastoma and its clinical relevance. Pathol. Oncol. Res. **20**(4), 777–787 (2014). https://doi.org/10.1007/s12253-014-9833-3
2. Hanif, F., Muzaffar, K., Perveen, K., Malhi, S.M., Simjee, S.U.: Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment. Asian Pac. J. Cancer Prev. APJCP 18(1), 3 (2017). https://doi.org/10.22034/APJCP.2017.18.1.3
3. Horbinski, C., Berger, T., Packer, R.J., Wen, P.Y.: Clinical implications of the 2021 edition of the WHO classification of central nervous system tumours. Nat. Rev. Neurol. **18**(9), 515–529 (2022). https://doi.org/10.1038/s41582-022-00679-w
4. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-International Conference on Neural Networks. vol. 4, pp. 1942–1948. IEEE (1995). https://doi.org/10.1109/ICNN.1995.488968
5. Mahajan, P., Uddin, S., Hajati, F., Moni, M.A.: Ensemble learning for disease prediction: a review. Healthcare **11**(12), 1808 (2023). https://doi.org/10.3390/healthcare11121808
6. Nematzadeh, S., Kiani, F., Torkamanian-Afshar, M., Aydin, N.: Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: a bioinformatics study on biomedical and biological cases. Comput. Biol. Chem. **97**, 107619 (2022). https://doi.org/10.1016/j.compbiolchem.2021.107619
7. Noviandy, T.R., Idroes, G.M., Hardi, I.: Integrating explainable artificial intelligence and light gradient boosting machine for glioma grading. Inf. Health **2**(1), 1–8 (2025). https://doi.org/10.1016/j.infoh.2024.12.001
8. Nssibi, M., Manita, G., Korbaa, O.: Advances in nature-inspired metaheuristic optimization for feature selection problem: a comprehensive survey. Comput. Sci. Rev. **49**, 100559 (2023). https://doi.org/10.1016/j.cosrev.2023.100559
9. Reel, P.S., Reel, S., Pearson, E., Trucco, E., Jefferson, E.: Using machine learning approaches for multi-omics data analysis: a review. Biotechnol. Adv. **49**, 107739 (2021). https://doi.org/10.1016/j.biotechadv.2021.107739
10. Tasci, E., Jagasia, S., Zhuge, Y., Camphausen, K., Krauze, A.V.: GradWise: a novel application of a rank-based weighted hybrid filter and embedded feature selection method for glioma grading with clinical and molecular characteristics. Cancers **15**(18), 4628 (2023). https://doi.org/10.3390/cancers15184628
11. Tasci, E., Zhuge, Y., Camphausen, K., Krauze, A.: Glioma grading clinical and mutation features. UCI Mach. Learn. Repository (2022). https://doi.org/10.24432/C5R62J
12. Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., Krauze, A.V.: Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. Int. J. Mol. Sci. **23**(22), 14155 (2022). https://doi.org/10.3390/ijms232214155
13. Tiwari, A., Chugh, A., Sharma, A.: Ensemble framework for cardiovascular disease prediction. Comput. Biol. Med. **146**, 105624 (2022). https://doi.org/10.1016/j.compbiomed.2022.105624
14. Van Thieu, N., Mirjalili, S.: MEALPY: an open-source library for latest meta-heuristic algorithms in python. J. Syst. Architect. **139**, 102871 (2023). https://doi.org/10.1016/j.sysarc.2023.102871
15. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997). https://doi.org/10.1109/4235.585893

16. Zhan, Y., Hao, Y., Wang, X., Guo, D.: Advances of artificial intelligence in clinical application and scientific research of neuro-oncology: Current knowledge and future perspectives. Crit. Rev. Oncol. Hematol. 104682 (2025). https://doi.org/10.1016/j.critrevonc.2025.104682
17. Zhao, Z., et al.: Chinese glioma genome atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. Genomics Proteomics Bioinform. **19**(1), 1–12 (2021). https://doi.org/10.1016/j.gpb.2020.10.005

Evaluating an Amazigh Speech Recognition System in a Noisy Environment Using MFDWC Features

Fadwa Abakarim[1(✉)] and Abdenbi Abenaou[2]

[1] ENSEIRB-MATMECA, Bordeaux INP, University of Bordeaux, 33400 Talence, France
fadwa.abakarim@bordeaux-inp.fr

[2] National School of Applied Sciences, Ibn Zohr University, 80000 Agadir, Morocco

Abstract. This work proposes an Automatic Speech Recognition (ASR) system for the Amazigh language, designed to operate in both calm and noisy environments. The system preprocesses speech using Zero-Crossing Rate (ZCR) and Short-Time Energy (STE) to remove silence and compress the signals. Features are then extracted using Mel-Frequency Discrete Wavelet Coefficients (MFDWC). Classification is performed using a Support Vector Machine (SVM) with a One-versus-One (OvO) approach for multiclass problems. A database of 63600 recordings was created, including 53 isolated units (33 letters, 10 digits, and 10 words) from 120 Moroccan speakers, 60 native Tachelhit speakers and 60 non-native speakers. Recordings were made in a calm environment (personal office) and a noisy environment (café). Experimental results show that the proposed MFDWC based system outperforms traditional MFCC and DWT approaches, achieving recognition rates of 90.37% in calm conditions and 73.55% in noisy conditions. These results demonstrate the effectiveness and robustness of the system, providing a valuable tool for supporting Amazigh language preservation and applications such as language learning, voice-based interfaces, and automatic transcription.

Keywords: Automatic Speech Recognition · Amazigh language · Zero-Crossing Rate · Short-Time Energy · Mel-Frequency Cepstral Coefficients · Discrete Wavelet Transform · Mel-Frequency Discrete Wavelet Coefficients · Support Vector Machine · Calm environment · Noisy environment

1 Introduction

The development of Automatic Speech Recognition (ASR) systems for the Amazigh language is important, particularly in response to the growing demand for e-learning tools in this language. Implementing speech-based interfaces opens new perspectives for the development of applications and voice-controlled systems in Amazigh, thereby addressing the expectations of Amazigh-speaking communities seeking to benefit from speech technologies.

However, the development and deployment of ASR systems in real-world conditions face several challenges. The primary difficulty lies in the presence of various types

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 209–227, 2027.
https://doi.org/10.1007/978-3-032-30494-0_14

of noise stemming from environmental characteristics, communication channels, and recording devices. Moreover, the system may be used by speakers who differ considerably in terms of gender, age, and native or non-native status, making it necessary to account for speech variability within a diverse user population. All these factors directly affect the performance and reliability of speech recognition systems.

The objective of this work is to develop an Amazigh ASR system capable of operating in noisy conditions. Our previous study [1] focused on designing an Amazigh ASR system under a normal environment. The proposed system relies on the Mel-Frequency Discrete Wavelet Coefficients (MFDWC) [1, 2] feature extraction method, which offers effective time-frequency localization and improved robustness in noisy acoustic conditions.

A common challenge in signal processing and machine learning is the lack of sufficient input data. Training and evaluating intelligent systems typically require large and representative datasets to achieve accurate and reliable results.

In this study, we enhanced the previous database used in our earlier work. The initial version contained 33 letters and 10 digits of the Amazigh language recorded in a normal environment [1]. The updated version now includes 53 isolated Amazigh units (33 letters, 10 digits, and 10 words), collected from 120 speakers, among whom 60 are native speakers from the Agadir region and 60 are non-native speakers. The complete dataset contains 63600 audio samples, recorded in both a noisy environment (a café) and a calm environment (personal office).

The remainder of this article is organized as follows: Sect. 2 describes the methods employed to develop the system, Sect. 3 presents the experimental results, and Sect. 4 concludes the study.

2 Applied Methods

In recent years, there has been a growing interest in the Amazigh language, which has led to research efforts aimed at developing speech recognition systems specifically tailored to this language [3–11]. These initiatives seek to improve the accessibility of speech recognition technologies for Amazigh speakers, while also contributing to the preservation and promotion of their cultural heritage.

Previous studies on the development of automatic speech recognition systems for Amazigh have primarily focused on Tarifit [3–8]. However, in the present research, our focus is on the Amazigh language spoken in the city of Agadir and its surroundings, namely Tachelhit.

In the field of Amazigh speech recognition, Mel-Frequency Cepstral Coefficients (MFCC) are the most used features [3–11]. However, they are sensitive to noise in adverse acoustic environments [12–15]. A key step in their extraction involves the use of the Discrete Cosine Transform (DCT).

The DCT basis vectors have similar temporal and frequency resolutions, meaning that all coefficients are affected by noise present within a specific frequency band [16]. This limitation reduces the effectiveness of the MFCC method in noisy environments. To address this issue, the use of the Discrete Wavelet Transform (DWT) can be advantageous due to its improved time-frequency localization [12–16].

Furthermore, available memory space is an important criterion when choosing a feature extraction method that achieves good accuracy with a limited feature vector size. The DWT method is a suitable option when storage capacity is constrained, whereas the MFCC method requires more storage space [13, 15].

Consequently, in many cases, the MFCC method is combined with other feature selection techniques aimed at reducing the dimensionality of the extracted feature vectors while maintaining good accuracy. Principal Component Analysis (PCA) is one such technique [15].

As part of the development of a new methodology for creating a noise-robust Amazigh automatic speech recognition system, we propose an approach combining MFCC and DWT methods. This approach, called Mel-Frequency Discrete Wavelet Coefficients (MFDWC) [1], provides good time-frequency localization and enhanced robustness in noisy environments.

This article proposes an Amazigh ASR system based on three main stages: preprocessing, feature extraction, and classification.

2.1 Preprocessing

The combination of Zero Crossing Rate (ZCR) and Short-Time Energy (STE) [17, 18] provides an effective method for detecting both speech activity and silence in a vocal signal.

Low ZCR values combined with high STE indicate the presence of speech, i.e., pronounced words. Using this criterion, it is possible to identify segments containing active speech.

Segments where ZCR is high and STE is low, corresponding to silence, are removed, resulting in purified speech signals that retain only the pronounced words while effectively eliminating silent periods.

Figure 1 illustrates the operation of the ZCR-STE combination.

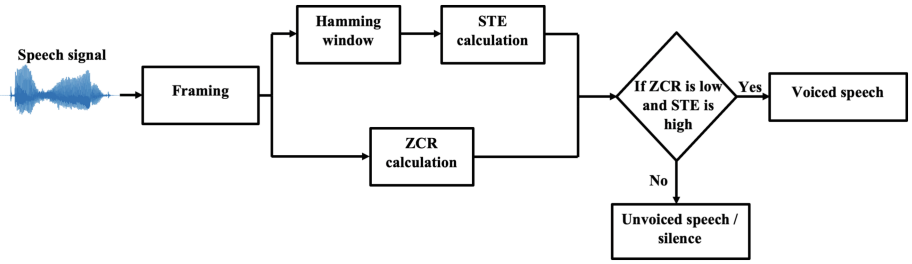


Fig. 1. The ZCR-STE process.

Figure 2 shows the speech signal of the Amazigh word “Azul” before and after applying the ZCR-STE combination.

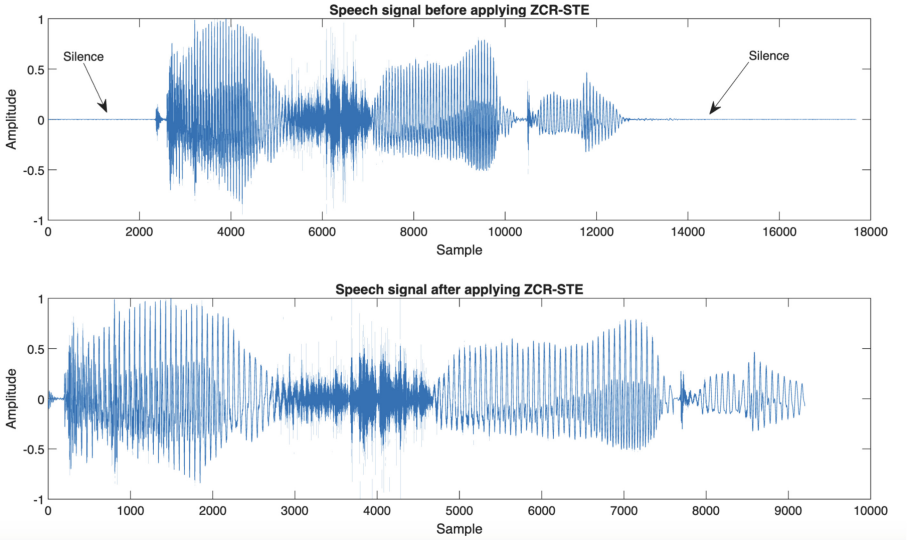


Fig. 2. Silence removal in the speech signal of the Amazigh word "Azul" using the ZCR-STE method.

2.2 Feature Extraction

After removing silent periods using the ZCR-STE combination, the speech signals are ready for feature extraction. In our previous work, we demonstrated the effectiveness of Mel-Frequency Discrete Wavelet Coefficients (MFDWC) for Amazigh speech recognition in a normal environment [1]. MFDWC combines the advantages of Mel-frequency analysis with DWT [14, 16], providing good time-frequency localization and improved robustness of extracted features.

In the present study, we extend the application of MFDWC to a noisy environment using a newly recorded dataset, where all speech samples were captured in both a café setting and a quiet environment, allowing for a comparison of results between clean and noisy conditions. This adaptation enables us to evaluate the robustness of MFDWC under real-world noise scenarios. While the overall feature extraction process follows the same principles as in our previous study [1], the performance under noisy conditions highlights the method's ability to maintain discriminative power despite challenging acoustic conditions.

Each speech frame is first processed through a high-pass pre-emphasis filter to enhance high-frequency components and reduce undesired low-frequency energy. The speech signal is then segmented into overlapping frames of 25 ms with a 10 ms overlap, and each frame is multiplied by a Hamming window to reduce spectral leakage. The Mel-scale filter bank is applied to the power spectrum of each frame, and the discrete wavelet transform is then performed on the logarithm of the filter bank energies to generate the MFDWC features [1].

The choice of the mother wavelet, Daubechies4, is consistent with our previous work [1]; however, unlike the previous study where 16 MFDWC features were obtained

in a normal environment, 11 coefficients are extracted in this study to optimize the representation for noisy conditions.

Figure 3 illustrates the complete process of MFDWC feature extraction.

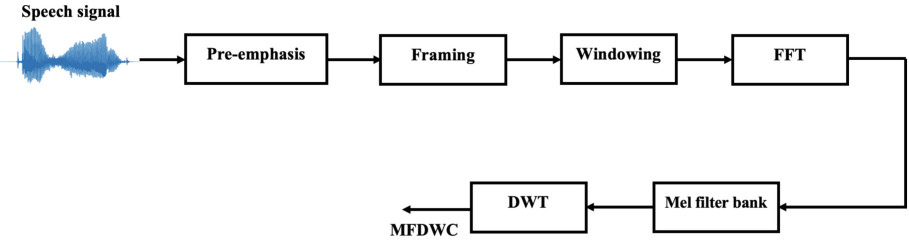


Fig. 3. The MFDWC feature extraction process [1].

2.3 Classification

After feature extraction, the next step is classification, which is crucial for achieving high recognition performance. Various models are commonly used in speech recognition, including supervised approaches such as Multilayer Perceptron (MLP) [19, 20] and Hidden Markov Models (HMM) [14, 21], as well as unsupervised methods such as Dynamic Time Warping (DTW) [22] and K-means [23]. Each method has specific limitations: MLPs may struggle with large vocabularies, HMMs rely on the Markov assumption that ignores long-term dependencies, DTW requires extensive computation and memory, and K-means may perform poorly when clusters are not well separated [13, 14].

Considering these limitations, we adopted Support Vector Machine (SVM) for classification in this study. SVMs are widely recognized in speech recognition due to their ability to handle non-linearly separable data through kernel functions, capturing complex relationships between acoustic features and speech classes [14, 24–26]. They also offer high accuracy, robustness to noisy data, and efficient decision-making, making them particularly suitable for real-world noisy environments.

As SVM is inherently a binary classifier, multiclass problems must be decomposed into multiple binary classification tasks. Two common approaches are One-versus-All (OvA) and One-versus-One (OvO). In OvA, each class is treated as positive against all other classes as negative, requiring K binary classifiers for K classes. In OvO, a binary classifier is trained for each pair of classes, resulting in $K(K-1)/2$ classifiers. While OvO involves more classifiers, each is trained on fewer samples and solves a simpler two-class problem, often leading to faster and more robust multiclass classification.

In this study, we employed the OvO approach with a Radial Basis Function (RBF) kernel. This combination balances computational efficiency and accuracy while effectively handling noisy speech signals.

3 Experimental Results and Discussion

3.1 Amazigh Speech Database

For training and testing our system, we created a dedicated database. In our previous work [1], we developed an Amazigh speech database containing 33 letters and 10 digits recorded in a normal environment. In this study, our goal is to evaluate the MFDWC method using both noisy and calm data. The newly created database consists of 53 isolated Amazigh units: 33 letters, 10 digits, and 10 isolated words.

A total of 120 Moroccan speakers participated in this study, including 60 native speakers from the Agadir region and 60 non-native speakers who spoke Amazigh-Tachelhit for the first time. The complete database contains 63600 speech recordings (53 units x 10 repetitions x 120 speakers).

Table 1 presents the 10 isolated words of the Amazigh-Tachelhit language along with their correspondences in English and Arabic. The 33 letters and 10 digits are presented in our previous work [1].

Table 1. The 10 isolated words of the Amazigh-Tachelhit language with their English and Arabic correspondences.

Tifinaghe-IRCAM [1]	English transcription	English correspondence	Arabic transcription	Arabic correspondence
ⵄⴰⵣⵓⵍ	Azul	Hello	أزول	مرحبًا
ⵜⴰⵏⵎⵎⵉⵔⵜ	Tanmmirt	Thank you	تانميرت	شكراً
ⵜⴼⵓⵍⴽⵉ	Tfulki	Beautiful	تفولكي	جميلة
ⵉⴷⴳⴰⵎ	Idgam	Yesterday	إضكام	أمس
ⵖⴰⵙⵙⴰⴷ	ɣassad	Today	غَسّاد	اليوم
ⵄⴰⵛⴽⵉⴷ	Achkid	Come	أشكيد	تعال
ⵖⵉⵍⴰⴷ	ɣilad	Now	غيلاد	الآن
ⵉⵎⵏⵙⵉ	Imnsi	Dinner	إمنسي	عشاء
ⵉⵎⴽⵍⵉ	Imkli	Lunch	إمكلي	غداء
ⵄⵖⵓⵊⵊⵔ	Aɣujjr	Breakfast	أغوجّر	فطور

The database used in this study is divided into two distinct parts. The first 31800 recordings were made in a quiet environment, specifically in a personal office, by 60 speakers. These recordings were captured under controlled conditions with minimal background noise, providing a calm acoustic context suitable for speech signal analysis.

The remaining 31800 recordings were made in a noisy environment, specifically in a café, by the other 60 speakers.

The noisy environment consisted of ambient acoustic conditions typical of a café setting, including overlapping background conversations, background music, and general crowd noise. No artificial noise was added or mixed in post-processing; all recordings were made directly in the natural café environment without noise filtering or enhancement. This ensures that the noise conditions reflect real-world deployment scenarios for speech recognition systems.

Figure 4 illustrates the speech signals of the word "Azul" in both environments, quiet and noisy, along with their corresponding spectrograms.

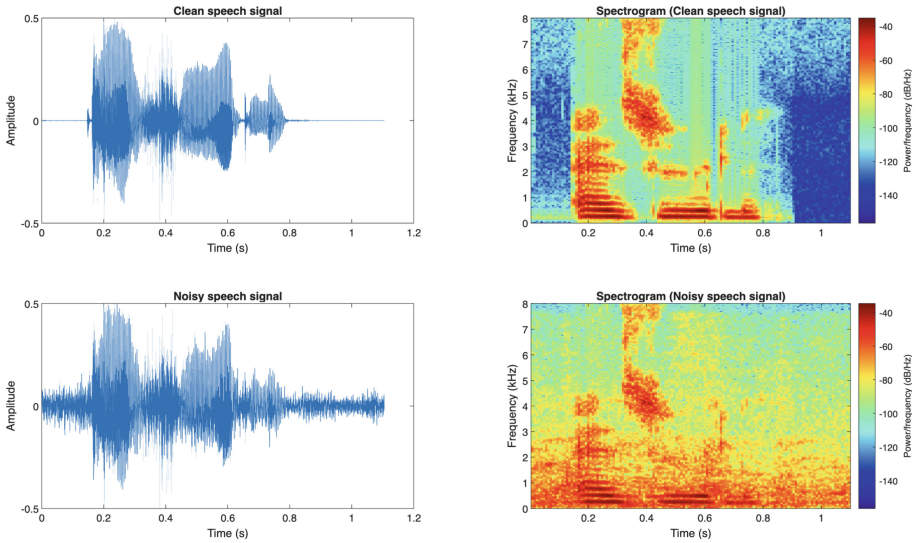


Fig. 4. The vocal signals of the word "Azul" in both quiet and noisy environments, along with their spectrograms.

A total of eight experiments were performed to train and test the system. The first four experiments used data from a quiet environment, while the last four experiments used data from a noisy environment.

For training, 70% of the database was used, while the remaining 30% was reserved for testing. Additionally, the speakers used in the testing phase were different from those used in the training phase.

Tables 2 and 3 summarize the different experiments conducted, along with the corresponding databases used for training and testing.

Table 2. The different experiments conducted in the calm environment.

Experiments	Calm environment			
	Training data		Test data	
	Number of audio files	Number of speakers	Number of audio files	Number of speakers
Experiment 1: 33 letters	13860	42 (30 adults and 12 children)	5940	18 (10 adults and 8 children)
Experiment 2: 10 digits (0–9)	4200	42 (30 adults and 12 children)	1800	18 (10 adults and 8 children)

(*continued*)

Table 2. (*continued*)

Experiments	Calm environment			
	Training data		Test data	
	Number of audio files	Number of speakers	Number of audio files	Number of speakers
Experiment 3: 10 isolated words	4200	42 (30 adults and 12 children)	1800	18 (10 adults and 8 children)
Experiment 4: 33 letters + 10 digits + 10 isolated words	22260	42 (30 adults and 12 children)	9540	18 (10 adults and 8 children)

Table 3. The different experiments conducted in the noisy environment.

Experiments	Noisy environment			
	Training data		Test data	
	Number of audio files	Number of speakers	Number of audio files	Number of speakers
Experiment 5: 33 letters	13860	42 (30 adults and 12 children)	5940	18 (10 adults and 8 children)
Experiment 6: 10 digits (0–9)	4200	42 (30 adults and 12 children)	1800	18 (10 adults and 8 children)
Experiment 7: 10 isolated words	4200	42 (30 adults and 12 children)	1800	18 (10 adults and 8 children)
Experiment 8: 33 letters + 10 digits + 10 isolated words	22260	42 (30 adults and 12 children)	9540	18 (10 adults and 8 children)

3.2 Calm Environment

Experimental Results: First Experiment. Table 4 reports the recognition performance obtained for the 33 Amazigh letters using 5940 speech recordings.

The MFCC based system achieved an overall accuracy of 78.27%, with the best results for YASH (ⵛ) and YAZ (ⵣ) (87.22% and 88.33%), while YA (ⴰ) and YO (ⵓ) showed lower accuracies (72.22% and 70.56%). The DWT based system yielded a higher overall accuracy of 83.01%, notably recognizing YAD (ⴷ) and YAM (ⵎ) with 91.11% and 92.78%, whereas YEY (ⵉ) and YI (ⵉ) remained comparatively low (75.00% and 73.89%). The combined MFDWC approach achieved the best performance, reaching an overall accuracy of 88.52%. It provided high recognition rates for YAD (ⴷ) and YAT (ⵟ)

(93.89% and 95.00%), although YA (ⴰ) and YI (ⵉ) remained among the least accurately recognized (83.33% and 81.67%). Overall, the MFDWC based system outperformed MFCC and DWT by 10.25% and 5.51%, respectively, confirming the effectiveness of the combined feature extraction strategy.

Table 4. Results of the first experiment.

Letters	Recognition rate (%)		
	MFCC	DWT	MFDWC
YA	72.22	77.78	83.33
YAB	75.00	81.11	85.00
YAG	77.78	79.44	90.00
YAGG	73.89	78.33	88.33
YAD	76.11	91.11	93.89
YADD	73.89	87.22	91.11
YEY	76.11	75.00	88.89
YAF	80.00	77.22	90.00
YAK	76.67	83.33	91.11
YAKK	73.33	85.56	87.22
YAH	82.78	88.33	89.44
YAHH	74.44	78.89	90.56
YAAA	76.11	81.11	87.78
YAKH	75.00	83.89	84.44
YAQQ	83.33	88.89	92.78
YI	76.11	73.89	81.67
YAJ	80.00	90.00	88.33
YAL	76.67	87.22	92.78
YAM	83.89	92.78	91.67
YAN	81.67	86.11	90.56
YO	70.56	80.00	84.44
YAR	82.78	84.44	90.00
YARR	83.89	81.67	88.33
YAGH	82.22	86.67	86.11
YAS	80.56	83.33	87.22
YASS	76.67	85.00	83.89
YASH	87.22	82.78	85.56
YAT	75.56	80.00	95.00
YATT	74.44	76.67	92.78

(*continued*)

Table 4. (*continued*)

Letters	Recognition rate (%)		
	MFCC	DWT	MFDWC
YAW	76.11	78.89	90.56
YAY	75.56	86.11	89.44
YAZ	88.33	83.89	85.00
YAZZ	83.89	82.78	83.89
Total recognition rate (%)	**78.27**	**83.01**	**88.52**

Experimental Results: Second Experiment. Table 5 summarizes the results of the second experiment, conducted on 1800 speech recordings to test the recognition performance of the first ten Amazigh digits.

The MFCC based system achieved an overall accuracy of 85.56%, with AMYA (ⴰⵎⵢⴰ) and KRAD (ⴽⵔⴰⴹ) obtaining the highest rates (91.67% and 93.33%), while SA (ⵙⴰ) and TZA (ⵜⵣⵉ) showed the lowest (80.00% and 78.89%). The DWT based system performed better, reaching 88.22% overall accuracy; YEN (ⵢⴻⵏ) and KRAD (ⴽⵔⴰⴹ) achieved the highest recognition rates (94.44% and 95.56%), whereas SA (ⵙⴰ) and TZA (ⵜⵣⵉ) remained the least accurate (81.67% and 82.78%). The combined MFDWC approach yielded the best performance, with an overall accuracy of 93.06%. YEN (ⵢⴻⵏ) and KRAD (ⴽⵔⴰⴹ) reached 96.11% and 97.78%, while KOZ (ⴽⵓⵣ) and SA (ⵙⴰ) recorded the lowest rates (90.56% and 89.44%). Overall, the MFDWC based system outperformed MFCC and DWT by 7.50% and 4.84%, respectively, confirming the effectiveness of combining both feature extraction techniques.

Table 5. Results of the second experiment.

Digits	Recognition rate (%)		
	MFCC	DWT	MFDWC
AMYA	91.67	90.56	93.33
YEN	88.33	94.44	96.11
SIN	83.89	88.33	91.11
KRAD	93.33	95.56	97.78
KOZ	81.67	83.33	90.56
SMMUS	83.89	84.44	91.67
SDES	85.56	90.00	93.89
SA	80.00	81.67	89.44
TAM	88.33	91.11	94.44
TZA	78.89	82.78	92.22

(*continued*)

Table 5. (*continued*)

Digits	Recognition rate (%)		
	MFCC	DWT	MFDWC
Recognition rate (%)	**85.56**	**88.22**	**93.06**

Experimental Results: Third Experiment. Table 6 summarizes the results of the third experiment, conducted on 1800 speech recordings to evaluate the recognition performance of ten isolated Amazigh-Tachelhit words.

The MFCC based system achieved an overall accuracy of 88.56%, with Tfulki (ⵜⴼⵓⵍⴽⵉ) and Idgam (ⵉⴷⴳⴰⵎ) obtaining the highest rates (94.44% and 95.56%), while Achkid (ⴰⵛⴽⵉⴷ) and Imnsi (ⵉⵎⵏⵙⵉ) recorded the lowest (83.33% and 81.67%). The DWT based system performed better, reaching 90.11% overall accuracy; Idgam (ⵉⴷⴳⴰⵎ) and Imkli (ⵉⵎⴽⵍⵉ) achieved the best recognition rates (96.11% and 95.00%), whereas ɣassad (ⵖⴰⵙⵙⴰⴷ) and Imnsi (ⵉⵎⵏⵙⵉ) remained the least accurate (85.56% and 83.89%). The combined MFDWC method yielded the highest overall accuracy of 94.67%. Tfulki (ⵜⵜⴼⵓⵍⴽⵉ) and Idgam (ⵉⴷⴳⴰⵎ) were particularly well recognized (97.78% and 98.89%), while ɣassad (ⵖⴰⵙⵙⴰⴷ) and Imnsi (ⵉⵎⵏⵙⵉ) showed the lowest rates (91.67% and 90.56%). Overall, the MFDWC based system surpassed MFCC and DWT by 6.11% and 4.56%, respectively, confirming the effectiveness of combining both feature extraction techniques for isolated Amazigh word recognition.

Table 6. Results of the third experiment.

Amazigh-Tachelhit words	Recognition rate (%)		
	MFCC	DWT	MFDWC
Azul	92.78	89.44	92.78
Tanmmirt	86.67	87.22	95.00
Tfulki	94.44	91.11	97.78
Idgam	95.56	96.11	98.89
ɣassad	88.33	85.56	91.67
Achkid	83.33	88.33	94.44
ɣilad	86.11	92.78	93.89
Imnsi	81.67	83.89	90.56
Imkli	88.89	95.00	96.11
Aghujjr	87.78	91.67	95.56
Recognition rate (%)	**88.56**	**90.11**	**94.67**

Tables 4, 5, and 6 clearly indicate that the recognition system for the 10 isolated Amazigh-Tachelhit words performs better than the recognition systems for the 33 letters

and 10 digits. This difference in performance can be largely attributed to the length of the units, as letters and digits are shorter compared to the tested Amazigh-Tachelhit words.

Experimental Results: Fourth Experiment. Table 7 presents the overall results of the fourth experiment, which used a total of 9540 speech recordings for testing. These recordings included a mixture of 33 letters, 10 digits, and 10 isolated Amazigh words.

The MFCC based recognition system for the 53 Amazigh units achieved an overall recognition rate of 81.29%. In comparison, the DWT based system reached an overall recognition rate of 85.18%, showing a difference of 3.89% and indicating that the DWT based system provides superior performance.

Furthermore, by combining the MFCC and DWT methods (MFDWC), the recognition system for the 53 Amazigh units achieved an overall recognition rate of 90.37%. This represents a significant improvement of 9.08% over the MFCC based system and a difference of 5.19% over the DWT based system.

These results clearly demonstrate that the combination of the two methods provides better recognition performance for Amazigh words.

Table 7. Results of the fourth experiment.

Overall data	Recognition rate (%)		
	MFCC	DWT	MFDWC
33 letters + 10 digits + 10 words	81.29	85.18	90.37

3.3 Noisy Environment

Experimental Results: Fifth Experiment. Table 8 summarizes the results of the fifth experiment, conducted on 5940 speech recordings to test the recognition performance of the 33 Amazigh letters under noisy conditions.

The MFCC based system obtained an overall accuracy of 49.73%, with YASH (ⵛ) and YAD (ⴷ) achieving the highest recognition rates (54.44% and 55.00%), while YA (ⵄ) and YO (ⵣ) were the least accurately recognized (45.00% and 43.89%). Compared to the calm environment, the accuracy dropped by 28.54%. The DWT based system performed better, reaching 66.84% overall accuracy; YAD (ⴷ) and YAM (ⵎ) achieved the highest rates (74.44% and 75.56%), whereas YA (ⵄ) and YI (ⵉ) recorded the lowest (60.00% and 58.33%). The decrease in accuracy between calm and noisy environments was 16.17%. When combining both techniques, the MFDWC method achieved the best performance, with an overall accuracy of 71.15%. YAD (ⴷ) and YAT (ⵟ) obtained the highest recognition rates (77.78% and 79.44%), while YA (ⵄ) and YI (ⵉ) remained the least recognized (65.56% and 63.89%). The drop in performance compared to calm conditions was 17.37%. Overall, the MFDWC based system outperformed MFCC and DWT by 21.42% and 4.31%, respectively, confirming the robustness and effectiveness of the combined feature extraction approach in noisy environments.

Table 8. Results of the fifth experiment.

Letters	Recognition rate (%)		
	MFCC	DWT	MFDWC
YA	45.00	60.00	65.56
YAB	48.33	68.33	71.67
YAG	51.11	72.78	69.44
YAGG	47.78	70.00	72.78
YAD	55.00	74.44	77.78
YADD	46.67	71.67	75.00
YEY	48.33	62.22	68.33
YAF	50.56	66.67	71.11
YAK	46.11	67.78	73.33
YAKK	47.78	65.00	72.22
YAH	47.22	66.67	68.89
YAHH	46.11	63.89	67.22
YAAA	51.67	71.11	72.78
YAKH	50.00	65.56	67.78
YAQQ	52.22	67.78	73.89
YI	48.33	58.33	63.89
YAJ	46.11	64.44	75.56
YAL	52.22	68.89	72.78
YAM	53.89	75.56	77.22
YAN	50.00	71.67	72.22
YO	43.89	66.67	66.67
YAR	47.78	70.00	72.78
YARR	48.89	63.89	68.89
YAGH	51.11	61.11	71.67
YAS	52.78	67.22	67.78
YASS	48.89	65.56	70.56
YASH	54.44	68.33	66.11
YAT	53.33	70.56	79.44
YATT	51.11	65.56	76.67
YAW	48.89	62.78	73.33
YAY	51.67	66.11	70.56
YAZ	53.33	63.89	67.22

(*continued*)

Table 8. (*continued*)

Letters	Recognition rate (%)		
	MFCC	DWT	MFDWC
YAZZ	50.56	61.11	66.67
Recognition rate (%)	**49.73**	**66.84**	**71.15**

Experimental Results: Sixth Experiment. Table 9 summarizes the results of the sixth experiment, conducted on 1800 speech recordings to evaluate the recognition of the first ten Amazigh digits under noisy conditions.

The MFCC based system achieved an overall accuracy of 56.50%, with AMYA (ⴰⵎⵢⴰ) and KRAD (ⴽⵔⴰⴷ) reaching the highest rates (66.67% and 68.33%). Conversely, TZA (ⵜⵣⴰ) and SA (ⵙⴰ) recorded the lowest accuracies (50.00% and 48.89%). Compared to calm conditions, performance decreased by 29.06%. The DWT based system performed significantly better, with an overall accuracy of 72.11%; YEN (ⵢⴻⵏ) and KRAD (ⴽⵔⴰⴷ) obtained the best results (76.67% and 78.33%), while TZA (ⵜⵣⴰ) and SA (ⵙⴰ) remained less accurately recognized (66.67% and 65.00%). The drop in accuracy relative to calm conditions was 16.11%. The combined MFDWC method further improved performance, achieving an overall accuracy of 78.83%. YEN (ⵢⴻⵏ) and KRAD (ⴽⵔⴰⴷ) showed the highest recognition rates (82.78% and 83.89%), whereas KOZ (ⴽⵓⵣ) and SA (ⵙⴰ) were the least recognized (76.11% and 75.00%). The accuracy difference between calm and noisy environments was reduced to 14.23%. Overall, the MFDWC based system outperformed MFCC and DWT by 22.33% and 6.72%, respectively, confirming the robustness and superiority of the combined feature extraction strategy under noisy conditions.

Table 9. Results of the sixth experiment.

Digits	Recognition rate (%)		
	MFCC	DWT	MFDWC
AMYA	66.67	72.22	79.44
YEN	58.33	76.67	82.78
SIN	53.88	71.11	78.33
KRAD	68.33	78.33	83.89
KOZ	50.56	70.00	76.11
SMMUS	52.22	71.67	77.22
SDES	55.00	73.89	80.00
SA	48.89	65.00	75.00
TAM	61.11	75.56	78.89
TZA	50.00	66.67	76.67

(*continued*)

Table 9. (*continued*)

Digits	Recognition rate (%)		
	MFCC	DWT	MFDWC
Recognition rate (%)	**56.50**	**72.11**	**78.83**

Experimental Results: Seventh Experiment. Table 10 summarizes the results of the seventh experiment, conducted on 1800 speech recordings to evaluate the recognition of ten isolated Amazigh-Tachelhit words in noisy conditions.

The MFCC based system achieved an overall accuracy of 61.94%, with Idgam (ⵉⴷⴳⴰⵎ) and Tfulki (ⵜⴼⵓⵍⴽⵉ) obtaining the highest rates (70.00% and 72.22%), while Achkid (ⴰⵛⴽⵉⴷ) and Imnsi (ⵉⵎⵏⵙⵉ) recorded the lowest (53.89% and 52.78%). The accuracy drop relative to calm conditions reached 26.62%. The DWT based system performed substantially better, achieving 75.33% overall accuracy; Imkli (ⵉⵎⴽⵍⵉ) and Idgam (ⵉⴷⴳⴰⵎ) reached the highest rates (78.89% and 80.56%), whereas Imnsi (ⵉⵎⵏⵙⵉ) and ɣassad (ⵖⴰⵙⵙⴰⴷ) were less accurately recognized (71.67% and 69.44%). The decrease in accuracy compared to calm conditions was 14.78%. The combined MFDWC method achieved the best performance, with an overall accuracy of 77.22%. Idgam (ⵉⴷⴳⴰⵎ) and Tfulki (ⵜⴼⵓⵍⴽⵉ) obtained the highest rates (82.22% and 83.33%), while ɣassad (ⵖⴰⵙⵙⴰⴷ) and Imnsi (ⵉⵎⵏⵙⵉ) showed the lowest (72.78% and 70.00%). The performance drop under noisy conditions was 17.45%. Overall, the MFDWC based system outperformed MFCC by 15.28% and DWT by 1.89%, confirming the robustness of the combined feature extraction strategy for isolated word recognition.

Table 10. Results of the seventh experiment.

Amazigh-Tachelhit words	Recognition rate (%)		
	MFCC	DWT	MFDWC
Azul	65.56	75.56	80.00
Tanmmirt	68.33	74.44	77.22
Tfulki	72.22	76.67	83.33
Idgam	70.00	80.56	82.22
ɣassad	55.56	69.44	72.78
Achkid	53.89	72.78	73.89
ɣilad	58.33	77.22	76.11
Imnsi	52.78	71.67	70.00
Imkli	66.67	78.89	75.00
Aghujjr	56.11	76.11	81.67
Recognition rate (%)	**61.94**	**75.33**	**77.22**

Tables 8, 9 and 10 clearly demonstrate that the system for recognizing the 10 isolated Amazigh-Tachelhit words outperforms the recognition systems for the 33 letters and 10 digits, which is also confirmed by Tables 4, 5, and 6 for the calm environment.

Experimental Results: Eighth Experiment. Table 11 summarizes the results of the eighth experiment, conducted on 9540 speech recordings combining 33 letters, 10 digits, and 10 isolated Amazigh words.

The MFCC based system achieved an overall accuracy of 51.43%, with a substantial drop of 29.86% between calm and noisy conditions. The DWT based system performed better, reaching 69.22% overall accuracy and a difference of 15.96% between environments. The combined MFDWC approach achieved the highest performance, with an overall accuracy of 73.55% and a difference of 16.82% between calm and noisy conditions. The MFDWC based system outperformed MFCC and DWT by 22.12% and 4.33%, respectively. Overall, these results confirm that the proposed MFDWC based system is the most effective and robust for Amazigh ASR across both calm and noisy environments, while the MFCC based system is notably more sensitive to noise.

Table 11. Results of the eighth experiment.

Overall data	Recognition rate (%)		
	MFCC	DWT	MFDWC
33 letters + 10 digits + 10 words	51.43	69.22	73.55

Although different speakers were used during the testing process compared to the training process, the system based on the proposed MFDWC method demonstrated strong performance. In a calm environment, it achieved an accuracy rate above 90.00%, while in a noisy environment, the accuracy remained higher than 73.00%.

Table 12 presents the results obtained by other systems in comparison with the proposed system.

Table 12. Results obtained by other systems and by the proposed system.

State of the art	Proposed system	Amazigh corpus	Number of speakers	Recognition rate (%)
[9]	CMU Sphinx based system	33 letters	34	90.00
[10]	CMU Sphinx based system	33 letters	9	88.00
[27]	MFCC and CNN based system	10 digits	42	91.75

(*continued*)

Table 12. (*continued*)

State of the art	Proposed system	Amazigh corpus	Number of speakers	Recognition rate (%)
[28]	CNN based system	18 isolated words	25	94.77
Current study	MFDWC based system	33 letters + 10 digits + 10 isolated words	120	90.37 (calm environment)
				73.55 (noisy environment)

4 Conclusion

This study proposes an Automatic Speech Recognition (ASR) system for the Amazigh language, built on an expanded database of 63600 recordings, including 33 letters, 10 digits, and 10 isolated words under quiet and noisy conditions. Preprocessing using ZCR and STE removed silent segments and compressed signals, improving efficiency. Features were extracted via a combination of MFCC and DWT (MFDWC), providing enhanced time-frequency representation and robustness to noise. Classification employed a One-versus-One SVM for effective multiclass separation. Experimental results demonstrate the system's superiority, achieving 90.37% accuracy in quiet conditions and 73.55% in noisy conditions, outperforming MFCC or DWT based methods and confirming the effectiveness of the proposed MFDWC architecture.

References

1. Abakarim, F., Abenaou, A.: Enhancing Amazigh speech recognition system with MFDWC-SVM. In: Gervasi, O., et al. (eds.) ICCSA 2023. LNCS, vol. 13956, pp. 471–488. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36805-9_31
2. Pandit, P., Makwana, P., Bhatt, S.: Automatic speech recognition of continuous speech signal of Gujarati language using machine learning. In: Sahni, M., Merigó, J.M., Jha, B.K., Verma, R. (eds.) Mathematical Modeling, Computational Intelligence Techniques and Renewable Energy. AISC, vol. 1287, pp. 147–159. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9953-8_13
3. Satori, H., Zealouk, O., Satori, K., ElHaoussi, F.: Voice comparison between smokers and non-smokers using HMM speech recognition system. Int. J. Speech Technol. **20**(4), 771–777 (2017). https://doi.org/10.1007/s10772-017-9442-0
4. El Ouahabi, S., Atounti, M., Bellouki, M.: Toward an automatic speech recognition system for Amazigh-Tarifit language. Int. J. Speech Technol. **22**(2), 421–432 (2019). https://doi.org/10.1007/s10772-019-09617-6
5. El Ouahabi, S., Atounti, M., Bellouki, M.: Optimal parameters selected for automatic recognition of spoken Amazigh digits and letters using Hidden Markov Model Toolkit. Int. J. Speech Technol. **23**(4), 861–871 (2020). https://doi.org/10.1007/s10772-020-09762-3

6. Telmem, M., Ghanou, Y.: The convolutional neural networks for Amazigh speech recognition system. Telkomnika (Telecommun. Comput. Electron. Control) **19**(2), 515–522 (2021). https://doi.org/10.12928/telkomnika.v19i2.16793
7. Elouahabi, S., Atounti, M., Bellouki, M.: Amazigh isolated-word speech recognition system using hidden Markov model toolkit (HTK). In: Proceedings of the 2016 International Conference on Information Technology for Organizations Development (IT4OD), pp. 1–7. IEEE (2016). https://doi.org/10.1109/IT4OD.2016.7479305
8. Satori, H., ElHaoussi, F.: Investigation Amazigh speech recognition using CMU tools. Int. J. Speech Technol. **17**(3), 235–243 (2014). https://doi.org/10.1007/s10772-014-9223-y
9. Telmem, M., Ghanou, Y.: Estimation of the optimal HMM parameters for Amazigh speech recognition system using CMU-Sphinx. Procedia Comput. Sci. **127**, 92–101 (2018). https://doi.org/10.1016/j.procs.2018.01.102
10. Telmem, M., Ghanou, Y.: Amazigh speech recognition system based on CMUSphinx. In: Ben Ahmed, M., Boudhir, A. (eds.) SCAMS 2017. LNNS, vol. 37, pp. 397–410. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74500-8_37
11. El Ghazi, A., Daoui, C., Idrissi, N., Fakir, M., Bouikhalene, B.: Système de reconnaissance automatique de la parole amazigh à base de la transcription en alphabet tifinagh. Revue Méditerranéenne des Télécommunications **1**(2), (2011)
12. Janse, P.V., Magre, S.B., Kurzekar, P.K., Deshmukh, R.R.: A comparative study between MFCC and DWT feature extraction technique. Int. J. Eng. Res. Technol. **3**(1), 3124–3127 (2014)
13. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. Multimed. Tools Appl. **80**(6), 9411–9457 (2021). https://doi.org/10.1007/s11042-020-10073-7
14. Cutajar, M., Gatt, E., Grech, I., Casha, O., Micallef, J.: Comparative study of automatic speech recognition techniques. In: Proceedings of the 2013 IET Signal Processing, vol. 7, no. 1, pp. 25–46 (2013). https://doi.org/10.1049/iet-spr.2012.0151
15. Labied, M., Belangour, A.: Automatic speech recognition features extraction techniques: a multi-criteria comparison. Int. J. Adv. Comput. Sci. Appl. **12**(8) (2021)
16. Tufekci, Z., Gowdy, J.N.: Feature extraction using discrete wavelet transform for speech recognition. In: Proceedings of the 2000 IEEE SoutheastCon. 'Preparing for The New Millennium' (Cat. No. 00CH37105), pp. 116–123. IEEE (2000). https://doi.org/10.1109/SECON.2000.845444
17. Shete, D. S., Patil, S. B., Patil, S.: Zero crossing rate and energy of the speech signal of Devanagari script. IOSR J. VLSI Signal Process. (IOSR-JVSP) **4**(1), 1–5 (2014)
18. Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B.D.: Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In: Proceedings of the 2008 American Society for Engineering Education (ASEE), pp. 1–7 (2008)
19. Ahad, A., Fayyaz, A., Mehmood, T.: Speech recognition using multilayer perceptron. In: Proceedings of IEEE Students Conference, pp. 103–109. IEEE (2002). https://doi.org/10.1109/ISCON.2002.1215948
20. Hennebert, J., Hasler, M., Dedieu, H.: Neural networks in speech recognition. Department of Electrical Engineering, Swiss Federal Institute of Technology (1994)
21. Alpaydin, E.: Introduction to Machine Learning. MIT Press, Cambridge (2020)
22. Lama, P., Namburu, M.: Speech recognition with dynamic time warping using MATLAB. CS 525, SPRING 2010-Project report (2010)
23. Majeed, S., Husain, H., Samad, S.A., Hussain, A.: Hierarchical k-means algorithm applied on isolated Malay digit speech recognition. In: Proceedings of International of Computer Science & Information Technology, pp. 33–37 (2012)

24. Padmanabhan, J., Johnson Premkumar, M.J.: Machine learning in automatic speech recognition: a survey. IETE Tech. Rev. **32**(4), 240–251 (2015). https://doi.org/10.1080/02564602.2015.1010611
25. Ghai, W., Singh, N.: Literature review on automatic speech recognition. Int. J. Comput. Appl. **41**(8), 42–50 (2012). https://doi.org/10.5120/5565-7646
26. Ali, H., Jianwei, A., Iqbal, K.: Automatic speech recognition of Urdu digits with optimal classification approach. Int. J. Comput. Appl. **118**(9), 1–5 (2015)
27. Boulal, H., Hamidi, M., Abarkan, M., Barkani, J.: Amazigh spoken digit recognition using a deep learning approach based on MFCC. Int. J. Electr. Comput. Eng. Syst. **14**(7), 791–798 (2023). https://doi.org/10.32985/ijeces.14.7.6
28. Daouad, M., Dadi, E.W.: Amazigh speech recognition using 1D CNN. In: Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security, pp. 1–6 (2023). https://doi.org/10.1145/3607720.3607760

A Heterogeneous Ensemble Meta-Model for Large-Scale Categorization of Portuguese Scientific Records in OasisBR

Sergio Antonio Andrade de Freitas[1(✉)], André Corrêa da Silva[2], Milena de Faria Silva[4], Renan Carneiro Batista[3], and Washington Luís Ribeiro de Carvalho Segundo[3]

[1] CEDIS, University of Brasília, Brasília, Brazil
sergiofreitas@unb.br
[2] Banco do Brasil, Brasília, Brazil
[3] IBICT - Brazilian Institute of Information in Science and Technology, Brasília, Brazil
{renanbatista,washingtonsegundo}@ibict.br
[4] DAC – UNICAMP, Campinas, Brazil
m299743@dac.unicamp.br

Abstract. Large multidisciplinary repositories require reliable topical categorization to support retrieval, analytics, curation, and downstream research evaluation. This paper revisits and reconstructs an industrial research effort aimed at categorizing approximately 1.6 million Portuguese scientific records from OasisBR according to the nine CAPES large research areas. The operational snapshot preserved two critical assets: a large pre-labeled subset of roughly 800,000 records and aggregated benchmark results for several machine-learning algorithms. Building on these materials, this paper formalizes the task as a large-scale scientific-record classification problem and proposes a heterogeneous ensemble meta-model that combines Random Forest, XGBoost, CatBoost, support vector classification, logistic regression, Gaussian naive Bayes, decision trees, and AdaBoost through performance-informed weighted voting. The reconstructed pipeline integrates metadata ingestion, lexical normalization, and feature consolidation, reducing the raw lexical space from approximately 200,000 surface forms to 3,140 predictive features. In the harmonized benchmark, Random Forest achieved the highest mean F1-score across the nine large areas (0.7073), followed closely by XGBoost (0.7042). At the area level, Random Forest was the best standalone learner in seven areas and XGBoost in two. More importantly, in the operational unseen-batch evaluation, the ensemble meta-model outperformed the best standalone baseline in all nine areas, increasing the average hit rate from 19.86% to 33.30%, an average gain of 13.43% points. The strongest gains were observed in Engineering (+18.08 points), Human Sciences (+17.04), and Biological Sciences (+15.91). The main contribution of this work is a reproducible computational fram-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 228–242, 2027.
https://doi.org/10.1007/978-3-032-30494-0_15

ing of large-scale scientific-record categorization for Portuguese metadata repositories, together with empirical evidence that heterogeneous ensembles offer more robust performance than single-model deployment in a difficult real-world setting.

Keywords: Text classification · Ensemble learning · Scientific databases · Information retrieval · Portuguese NLP · OasisBR

1 Introduction

The growth of scientific repositories has intensified the need for reliable large-scale categorization pipelines. In digital-library environments, categorization is not merely a convenience layer for browsing. It affects search quality, visibility of scientific production, statistical analyses, thematic monitoring, and the quality of downstream metadata services. Automated text categorization become a problem at the intersection of information retrieval, natural language processing, and scientific-database engineering [11,15,16].

This problem is pertinent in the Brazilian context. OasisBR, maintained by the Brazilian Institute of Information in Science and Technology, brings together open-access scientific production and research data from journals, institutional repositories, and digital libraries, thereby constituting an access point to articles, theses, dissertations, conference papers, datasets, preprints, and other scholarly objects [8]. Because the portal aggregates metadata from different institutions, record quality and topical consistency constitute operational requirements for the portal. Studies on metadata standardization in OasisBR indicate that interoperability and record normalization constitute conditions for information visibility and access in the portal [12].

A second domain requirement concerns the classification system adopted by CAPES. The CAPES table organizes knowledge into nine broad areas and subordinate hierarchical levels, providing a taxonomy used in research and graduate-education evaluation processes [4]. In repositories oriented toward scientific information retrieval and policy support, aligning metadata with this taxonomy favors record aggregation, thematic analysis, and the identification of related scholarly production.

This article addresses the problem of automatically categorizing OasisBR records into CAPES broad areas by means of machine learning. The project that originated this study considered a corpus of approximately 1.6 million records, of which about 800,000 already contained assigned categories and were used as training evidence. Although the initial version of the study presented empirical results, it did not formalize the computational workflow, the baseline definition, or the meta-model logic at the level required by contemporary LNCS papers. The present article reconstructs the study as a computational contribution grounded in reproducible procedures.

The central contribution consists in reformulating the generic notion of a "meta-model" as a heterogeneous ensemble meta-model. Instead of relying on

a single classifier, the proposed system combines distinct base models whose performance varies according to the lexical profiles of the records and the research areas considered. This design is motivated by the well-known observation that scientific texts exhibit heterogeneous distributions, domain-specific jargon, and area-dependent class boundaries, which makes single-model deployment brittle in practice [1,9,10,15]. Ensemble learning is a natural candidate for this setting because it can integrate complementary decision patterns while reducing the risk of over-specialization [2,6,14].

The contributions of this paper are fourfold:

1. it reframes the OasisBR categorization task as a large-scale scientific-record classification problem grounded in information-retrieval and scientific-database requirements;
2. it formalizes a heterogeneous ensemble meta-model based on performance-informed weighted voting over eight machine-learning learners;
3. it consolidates the preserved benchmark evidence into a coherent two-stage evaluation composed of a harmonized model-selection benchmark and an operational unseen-batch comparison; and
4. it identifies practical lessons for large multilingual or Portuguese-language repositories, especially regarding lexical consolidation, model robustness, and deployment-oriented evaluation.

The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 presents the problem setting and dataset context. Section 4 describes the proposed methodology. Section 5 reports the results. Section 6 discusses implications, limitations, and future work. Threats to validity are summarized in Sect. 7, and Sect. 8 concludes the paper.

2 Related Work

2.1 Scientific-Document and Research-Article Classification

Automated text categorization has a long research tradition, with classical surveys already noting the importance of representation choices, feature selection, and algorithmic bias toward particular data distributions [15,16]. More recent work has specialized this agenda toward scientific and scholarly corpora. Lu et al. proposed an ensemble approach for identifying artificial-intelligence articles in the Web of Science, combining rule-based filtering, SciBERT, decision trees, and support vector machines to improve recall and F1-score over keyword-only retrieval [10]. Lee et al. used pre-trained BERT to classify scientific articles about electromagnetic-field health risks into research categories, showing that title-and-abstract classification can support the curation of specialized scientific corpora [9]. These studies confirm two points that are central to the present work: first, scientific texts can be classified effectively from metadata fields such as title and abstract; second, ensemble or hybrid approaches are especially attractive in research-domain settings with fuzzy boundaries.

Related evidence also appears outside scholarly metadata repositories. Aubaid et al. compared machine-learning and rule-based embedding techniques for document classification and highlighted the role of representation and preprocessing choices in final performance [1]. Rahman et al. evaluated several ensemble configurations for comment classification and showed that heterogeneous combinations of learners can outperform homogeneous or single-model baselines [14].

2.2 Ensemble Learning for Text Classification

In machine learning, ensemble methods combine predictors with the objective of increasing generalization capacity and reducing sensitivity to variation in the data. Bagging and boosting constitute reference strategies in this literature [2,6]. Random Forest combines decision trees estimated from data resampling and from attribute subsets selected through randomization, a procedure suited to representations with many dimensions and sparsity [2]. Families of gradient boosting, such as XGBoost and CatBoost, extend the logic of boosting through training with regularization and support for scalability, with applications in tabular and textual representations [3,13]. Support vector classifiers retain relevance in high-dimensional feature spaces because of their margin-based formulation [5]. Logistic regression, in turn, provides linear decision boundaries and probability estimates in text-classification tasks [7].

The literature on text classification indicates that no learner achieves superior performance across all corpora and all label structures [15]. This limitation is relevant in tasks involving records of scholarly production, intersections among fields of knowledge, distinct metadata sources, and terminological variation. In such settings, ensemble learning can mitigate the fragility of individual models through the aggregation of decision patterns. The methodological problem consists of selecting base learners and defining a combination rule capable of preserving stability during system use.

2.3 Repository Engineering and Metadata Context

OasisBR is not a document collection, but a federated service for scientific information that aggregates objects from journals, repositories, and digital libraries [8]. In this type of infrastructure, classification quality depends on the quality of the available metadata. Marques et al. argue that metadata standardization constitutes a requirement for dissemination, interoperability, and access in the OasisBR portal [12]. This finding has consequences for machine learning, since category assignment depends on the consistency of fields such as title, abstract, keywords, and language tagging.

The CAPES taxonomy introduces an additional constraint to the problem. Because it is used to organize the evaluation of research and graduate education in Brazil, associating repository records with CAPES areas can support reporting, indexing, thematic analysis, and retrieval of related scholarly production [4]. The present paper therefore sits at the interface of machine learning and scientific-database engineering rather than in pure algorithm design.

3 Problem Formulation and Dataset Context

3.1 Task Definition

The task addressed in this paper is the automatic categorization of OasisBR scientific records into the nine CAPES large areas: Agricultural Sciences, Biological Sciences, Health Sciences, Exact and Earth Sciences, Human Sciences, Applied Social Sciences, Engineering, Linguistics, Literature and Arts, and Multidisciplinary. The article is based on a project snapshot in which the working repository contained approximately 1.6 million records, with roughly half of them already categorized and available as labeled evidence.

At a practical level, the input for each record is composed of metadata fields that are common in scientific repositories, especially title, abstract, and keywords. The output is a CAPES large-area label. From an information-systems perspective, the objective is to support retrieval, repository curation, and large-scale metadata enrichment. From a machine-learning perspective, the challenge lies in dealing with Portuguese scientific language, heterogeneous terminology, and variable metadata quality across areas.

3.2 Preserved Project Evidence and Reconstruction Strategy

The project materials preserved three classes of evidence: (i) a narrative description of the pipeline, (ii) aggregated benchmark indicators for multiple algorithms across the nine large areas, and (iii) an operational comparison between a single-model baseline and an ensemble meta-model on unseen document batches. However, the archived materials did not preserve all experiment scripts, fold-level traces, or a fully specified reproducibility package. To maximize scientific clarity while remaining faithful to the preserved evidence, this paper reconstructs the study around a conservative two-stage evaluation:

1. a harmonized benchmark used for model comparison and baseline selection across the nine large areas; and
2. an operational unseen-batch evaluation used to compare the selected standalone baseline against the proposed ensemble meta-model.

This reconstruction strategy avoids overclaiming unavailable details while still providing a coherent and reproducible computational framing.

3.3 The Feature Space

In revisiting the materials, we retain the observation that natural-language preprocessing and feature consolidation reduced the raw lexical space from approximately 200,000 surface forms to 3,140 predictive features. Although the historical version of the study does not specify all preprocessing procedures in detail, we interpret this reduction as consistent with a computational pipeline comprising metadata normalization, token filtering, stopword removal, and vocabulary consolidation. For Portuguese scientific corpora, such consolidation is methodologically important, since inflectional variation, multiword terminology, and metadata heterogeneity tend to increase representational sparsity.

4 Methodology

4.1 Overview

Figure 1 summarizes the proposed pipeline. The process begins with metadata ingestion from OasisBR records. The title, abstract, and keywords are normalized and consolidated into a predictive representation. Multiple base learners are then trained and evaluated. Finally, their outputs are integrated by a performance-informed ensemble meta-model.

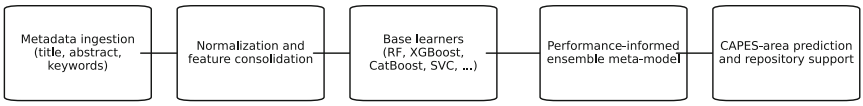


Fig. 1. Overview of the reconstructed categorization pipeline.

4.2 Metadata Preprocessing and Feature Engineering

The preprocessing stage is designed to stabilize the lexical representation of Portuguese scientific metadata. The preserved project description indicates that the final predictive space was obtained after aggressive lexical consolidation, reducing the raw space to 3,140 features. In the reconstructed formulation, the preprocessing stage comprises the following operations:

1. lowercasing and punctuation normalization for title, abstract, and keyword fields;
2. removal of obviously non-informative symbols and duplicated lexical fragments;
3. Portuguese stopword filtering and suppression of low-information tokens;
4. consolidation of morphological and near-duplicate variants into a reduced predictive vocabulary; and
5. document-vector construction over the consolidated lexical space.

The purpose of this stage is not only dimensionality reduction. It also reduces noise caused by repository heterogeneity. In scientific databases, records often originate from different institutional practices and export conventions. Vocabulary consolidation therefore acts as both a machine-learning step and a metadata-engineering step.

4.3 Base Learners

Eight base learners are considered in the harmonized benchmark: AdaBoost, CatBoost, Decision Tree, Gaussian Naive Bayes, Logistic Regression, Random Forest, Support Vector Classification (SVC), and XGBoost. These learners were selected because they cover distinct inductive biases and model families:

- tree ensembles and boosting methods are effective when complex non-linear decision patterns exist [2,3,6,13];
- margin-based classifiers are traditionally competitive in sparse text spaces [5]; and
- linear probabilistic models provide a stable reference point for high-dimensional text categorization [7,15].

The intent is not to claim that these eight learners exhaust the search space. Rather, they represent a sufficiently heterogeneous committee for robust ensemble construction.

4.4 Standalone Baseline Selection

The preserved benchmark reports per-area accuracy, precision, recall, and F1-score for several algorithms. From these results, the best standalone learner is selected for each large area according to F1-score. Table 1 shows the selected area-wise baselines.

Table 1. Best standalone learner by CAPES large area in the harmonized benchmark

Large area	Best model	Acc.	Prec.	Rec.	F1
Agricultural Sciences	Random Forest	0.679	0.821	0.679	0.706
Biological Sciences	Random Forest	0.654	0.852	0.654	0.709
Health Sciences	Random Forest	0.652	0.840	0.652	0.697
Exact and Earth Sciences	XGBoost	0.654	0.826	0.654	0.694
Human Sciences	XGBoost	0.670	0.813	0.670	0.705
Applied Social Sciences	Random Forest	0.696	0.897	0.696	0.753
Engineering	Random Forest	0.623	0.832	0.623	0.678
Linguistics, Literature and Arts	Random Forest	0.712	0.772	0.712	0.716
Multidisciplinary	Random Forest	0.700	0.838	0.700	0.721
Best-model frequency		Random Forest in 7 areas; XGBoost in 2 areas			

This area-wise best learner forms the simple-model baseline later used in the operational comparison.

4.5 Performance-Informed Ensemble Meta-Model

The reconstructed meta-model aggregates the class-probability outputs of all base learners through weighted voting. Let $p_m(c \mid d)$ denote the probability assigned by learner m to class c for document d. We define a global performance weight w_m from the harmonized benchmark as

$$w_m = \frac{\overline{F1}_m}{\sum_{j=1}^{M} \overline{F1}_j}, \tag{1}$$

where $\overline{F1}_m$ is the mean F1-score of learner m across the nine large areas. The meta-model score is then

$$S(c \mid d) = \sum_{m=1}^{M} w_m \, p_m(c \mid d), \tag{2}$$

and the final prediction is

$$\hat{y}(d) = \arg\max_{c} S(c \mid d). \tag{3}$$

This design is intentionally conservative. It does not require training an additional black-box stacker, yet it still exploits the relative strengths of stronger learners. Table 2 reports the mean benchmark performance used to derive the weights.

Table 2. Mean benchmark performance by algorithm across the nine large areas

Algorithm	Acc.	Prec.	Rec.	F1
Random Forest	0.670	0.833	0.670	0.707
XGBoost	0.667	0.826	0.667	0.704
CatBoost	0.660	0.816	0.660	0.696
SVC	0.659	0.822	0.659	0.696
Logistic Regression	0.638	0.790	0.638	0.670
GaussianNB	0.503	0.647	0.503	0.499
Decision Tree	0.485	0.581	0.485	0.499
AdaBoost	0.469	0.596	0.469	0.485

The weight distribution is therefore naturally dominated by Random Forest and XGBoost, but the ensemble still benefits from the complementary behavior of CatBoost, SVC, and Logistic Regression.

4.6 Evaluation Protocol

We organized the evaluation in two settings.

Harmonized benchmark. The first setting uses the preserved labeled evidence to compare base learners and select the area-wise best standalone baseline. The reported metrics are accuracy, precision, recall, and F1-score.

Operational unseen-batch evaluation. The second setting compares the simple-model baseline with the proposed meta-model on batches of previously unseen documents sampled from each large area. We treat this setting as a stricter and more deployment-oriented evaluation scenario than the harmonized benchmark. It therefore approximates the conditions faced by a repository service

when assigning categories to newly ingested records. At this stage, we retain two indicators: the number of exact hits and the hit rate for each broad area.

We adopt this two-stage protocol to support a more precise interpretation of the results. Benchmarks based on labeled and curated subsets allow us to estimate the relative quality of the models under controlled conditions. Operational batches, however, tend to exhibit greater noise, distribution shift, metadata sparsity, and topical ambiguity. Thus, a model that achieves competitive F1 performance in the benchmark may not preserve the same behavior under deployment conditions. We motivate the proposed meta-model from this gap between experimental performance and operational performance.

5 Results

5.1 Benchmark for Standalone Learners

In the harmonized benchmark, we observe a clear performance hierarchy among the base learners. Random Forest achieved the highest mean F1-score across the nine large areas (0.707), followed by XGBoost (0.704), CatBoost (0.696), and SVC (0.696). Logistic Regression remained competitive, although below this leading group. Gaussian Naive Bayes, Decision Tree, and AdaBoost formed a lower-performing cluster, with mean F1-scores close to or below 0.50. Figure 2 summarizes this ranking.

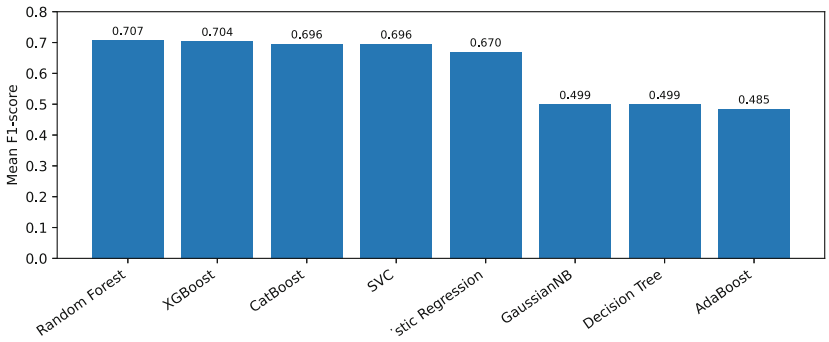


Fig. 2. Mean F1-score by algorithm in the harmonized benchmark.

We noted that two points deserve emphasis. First, tree-based ensembles dominate the ranking, which is consistent with their capacity to model non-linear feature interactions and to handle heterogeneous document distributions. Second, the small performance gap between Random Forest and XGBoost suggests that both models capture relevant, but not identical, patterns in the data. This result motivates their combination in an ensemble, rather than the selection of a single classifier.

At the area level, Random Forest was the best standalone learner in seven of the nine large areas, while XGBoost led in Exact and Earth Sciences and Human Sciences. The highest area-specific F1-score was observed in Applied Social Sciences (0.753 with Random Forest), followed by Multidisciplinary (0.721) and Linguistics, Literature and Arts (0.716). The lowest value among the selected best models was found in Engineering (0.678), suggesting that this area may be more lexically heterogeneous or more difficult to separate using metadata-only inputs.

5.2 Operational Comparison Between the Simple Model and the Meta-Model

Table 3 reports the unseen-batch comparison between the selected standalone baseline and the ensemble meta-model. The meta-model improved the hit rate in all nine large areas.

Table 3. Operational unseen-batch comparison between the simple-model baseline and the proposed meta-model

Large area	N	Simple hits (%)	Meta hits (%)	Gain (pts)
Agricultural Sciences	1497	307 (20.51)	432 (28.86)	+8.35
Biological Sciences	2495	235 (9.42)	632 (25.33)	+15.91
Health Sciences	1497	313 (20.91)	528 (35.27)	+14.36
Exact and Earth Sciences	1996	515 (25.80)	796 (39.88)	+14.08
Human Sciences	1497	315 (21.04)	570 (38.08)	+17.04
Applied Social Sciences	1497	314 (20.98)	488 (32.60)	+11.62
Engineering	1996	362 (18.14)	723 (36.22)	+18.08
Linguistics, Literature and Arts	1996	595 (29.81)	738 (36.97)	+7.16
Multidisciplinary	1497	182 (12.16)	396 (26.45)	+14.29
Average	–	**19.86**	**33.30**	**+13.43**

Figure 3 shows the same pattern graphically. The gains are consistent and substantial.

We observed the largest gains in Engineering (+18.08 %), Human Sciences (+17.04 %), and Biological Sciences (+15.91%). These areas appear to benefit more from ensemble aggregation, as they encompass heterogeneous terminologies and partially overlapping vocabularies. By contrast, Linguistics, Literature, and Arts showed the highest baseline hit rate and the smallest gain, suggesting that its lexical signal may be more homogeneous or more stable.

5.3 Benchmark Performance and Operational Accuracy

Area-level F1-scores in the range of 0.68–0.75 may seem difficult to reconcile with operational hit rates between 25% and 40%. This difference reflects the

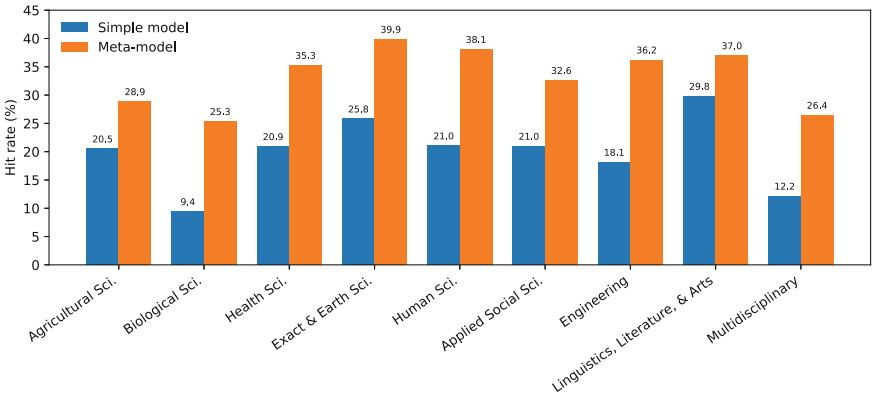


Fig. 3. Operational hit-rate comparison on unseen batches.

evaluation setting rather than an inconsistency in the results. The harmonized benchmark captures model behavior on curated labeled data, whereas the operational batches approximate deployment conditions. Records ingested into the repository may contain ambiguous metadata, limited abstracts, noisy keywords, or terminology associated with emerging research topics. Lower exact-hit rates are therefore expected in this setting.

This distinction is methodologically relevant for repository engineering. Evaluation in this domain should not be restricted to curated benchmark scores. Benchmarking supports the identification of candidate learners, but operational evaluation is required to assess whether the classification logic remains stable under input variability. The proposed meta-model addresses this requirement by combining models with complementary decision patterns and by reducing sensitivity to instability in the operational data.

6 Discussion

6.1 The Predominance of Tree-Based Ensembles

We interpret the predominance of Random Forest and XGBoost as consistent with the representational structure of scientific metadata. Titles, abstracts, and keywords encode lexical evidence together with interactions that vary across research areas and are not adequately captured by a strictly linear decision function. Tree-based ensembles are therefore appropriate for this setting, since they can model feature interactions without imposing an assumption of linear separability [2,3]. They also provide a suitable inductive bias for heterogeneous feature spaces, in which different subsets of terms may be informative for different areas of knowledge.

We relate this result to the heterogeneity of the predictive signal. Some records contain terminology that is strongly associated with a specific CAPES area, which makes their classification less ambiguous. Other records, however,

are shorter, interdisciplinary, weakly normalized, or semantically ambiguous. In this context, tree-based ensembles can reduce dependence on a single partition of the feature space. Random Forest does so by aggregating decisions from multiple trees estimated over resampled data and feature subsets, whereas XGBoost does so by sequentially fitting additive trees that correct residual errors. This helps explain why both methods performed better than simpler baseline models in the harmonized benchmark.

6.2 Area-Specific Gains of the Meta-Model

6.3 Area-Specific Gains of the Meta-Model

We observed the largest gains in Engineering, Human Sciences, and Biological Sciences. Although these areas differ in substantive scope, they share a common source of classification difficulty: internal lexical heterogeneity. Engineering records may contain specialized terminology distributed across several subfields. Human Sciences records often involve conceptual vocabularies and interdisciplinary intersections. Biological Sciences combine domain-specific terminology with expressions associated with health, environmental, and life-science topics. Under these conditions, we do not expect a single learner to provide uniform reliability across all lexical configurations. We address this limitation through a meta-model that combines classifiers with complementary inductive biases.

We interpret these gains as evidence that deployment-oriented categorization in large scientific repositories should prioritize robustness rather than optimization restricted to curated validation settings. From the perspective of repository operation, marginal improvements on a controlled validation subset are less consequential than reducing classification failures when new records are ingested. For this reason, we consider the average improvement of 13.43 % in operational hit rate to be one of the most relevant empirical findings of this study.

6.4 Implications for Scientific Databases and Information Retrieval

From the perspective of scientific-database engineering, our results indicate that topical classification should be understood as a repository-service layer rather than as an isolated machine-learning task. In large-scale repository infrastructures, improved categorization can enhance search filters, support thematic browsing, and provide more consistent analytical indicators for repository management.

The results also reinforce the methodological role of preprocessing. In federated repositories such as OasisBR, lexical normalization and metadata consolidation are not secondary implementation steps. They condition the quality of the feature space and directly affect the robustness of the classification models. Consequently, the performance of the proposed meta-model depends not only on the learning algorithms, but also on the capacity of the pipeline to reduce noise, sparsity, and inconsistency in the metadata.

A further implication concerns evaluation design. Scientific repositories continuously ingest records from multiple institutions and disciplinary communities,

which exposes classification systems to metadata drift and input heterogeneity. Models that perform well under static benchmark conditions may lose reliability in operational use. For this reason, we argue that repository-oriented evaluation should incorporate deployment-oriented scenarios. The two-stage protocol adopted in this study provides a pragmatic mechanism for assessing both controlled benchmark performance and operational behavior under more variable input conditions.

6.5 Future Methodological Extensions

The preserved project notes indicate a promising future direction based on synonym-aware feature engineering for Portuguese scientific text. Such an extension could reduce sparsity further by clustering semantically related surface forms, especially across institution-specific wording variants. Because this idea was not preserved under the harmonized benchmark, it is treated here as future work rather than as part of the evaluated contribution.

A second extension concerns modern transformer-based architectures. The related literature shows that BERT-family models can perform strongly on scientific article classification [9,10]. A natural continuation of this work is therefore a hybrid architecture in which transformer encoders provide semantic embeddings and the current ensemble acts either as a calibration or ranking layer.

7 Threats to Validity

Internal validity. The main threat concerns the archival nature of the preserved materials. Not all fold-level traces, hyperparameter logs, or operational scripts were retained. To mitigate this threat, the paper explicitly distinguishes between preserved evidence and reconstructed formalization, and it avoids unsupported causal claims.

Construct validity. The harmonized benchmark uses accuracy, precision, recall, and F1-score, while the operational stage reports exact-hit rates on unseen batches. These metrics capture related but not identical aspects of performance. We therefore interpret them as complementary rather than directly interchangeable.

External validity. The study focuses on Portuguese scientific metadata and on the CAPES taxonomy. Although the computational strategy is portable, the numerical results may not transfer directly to repositories with different metadata conventions, languages, or taxonomic systems.

Conclusion validity. The meta-model improved the operational hit rate in all nine areas, which is a strong empirical pattern. However, stronger statistical analysis would require fold-level or document-level prediction traces. The present paper therefore reports a robust comparative tendency rather than inferential significance testing.

8 Conclusion and Future Work

This paper reconstructed and strengthened a large-scale scientific-record categorization study for OasisBR and positioned it as a contribution to scientific databases, information retrieval, and applied machine learning. Using approximately 800,000 pre-labeled records from a working repository snapshot of about 1.6 million items, we formalized a heterogeneous ensemble meta-model for CAPES large-area categorization of Portuguese scientific metadata.

The empirical evidence supports three main conclusions. First, among the standalone learners, Random Forest and XGBoost are the most reliable choices, with Random Forest achieving the highest mean F1-score and dominating seven of the nine area-wise benchmarks. Second, a performance-informed ensemble meta-model improves robustness substantially in the harder operational setting, increasing the average hit rate from 19.86% to 33.30%. Third, metadata normalization and feature consolidation are decisive components of the pipeline, not marginal preprocessing details.

For repository operators, the results indicate that large-scale categorization of scientific records should combine heterogeneous ensemble methods, deployment-oriented evaluation protocols, and systematic metadata engineering. These components are not ancillary to the classification task; rather, they constitute conditions for the reliability of repository-scale categorization.

The study suggests some directions for further work. These include synonym-aware feature engineering, transformer-based embeddings for Portuguese scientific text, hierarchical classification beyond the CAPES large-area level, and continuous-learning pipelines capable of accommodating changes in repository content over time.

Acknowledgments. The authors acknowledge the project team involved in the original OasisBR categorization effort and the institutional collaborators who contributed to preserving the benchmark indicators used in this reconstruction. The authors also acknowledge the financial support provided by the Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF) through a research project.

References

1. Aubaid, A.M., Mishra, A., Mishra, A.: Machine learning and rule-based embedding techniques for classifying text documents. Int. J. Syst. Assur. Eng. Manag. **15**, 5637–5652 (2024). https://doi.org/10.1007/s13198-024-02555-w
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
3. 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). https://doi.org/10.1145/2939672.2939785

4. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior: Tabela de Áreas de conhecimento/avaliação. CAPES official website (2024). Updated 2024-02-01. https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/instrumentos/documentos-de-apoio/tabela-de-areas-de-conhecimento-avaliacao. Accessed 07 Mar 2026
5. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). https://doi.org/10.1006/jcss.1997.1504
7. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley, 3 edn. (2013). https://doi.org/10.1002/9781118548387
8. Instituto Brasileiro de Informação em Ciência e Tecnologia: O oasisbr. Oasisbr official website (2025). https://www.oasisbr.ibict.br/vufind/about/home. Accessed 07 Mar 2026
9. Lee, S.W., Kwon, J.H., Kim, D., Kim, E.J.: Research category classification of scientific articles on human health risks of electromagnetic fields using pre-trained BERT. ICT Express **10**(2), 336–341 (2024). https://doi.org/10.1016/j.icte.2023.08.006
10. Lu, M., Tang, L., Zhou, X.: An ensemble approach for research article classification: a case study in artificial intelligence. PeerJ Comput. Sci. **10**, e2521 (2024). https://doi.org/10.7717/peerj-cs.2521
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
12. Marques, G.S., Santos, B.E.D., Lima, M.C.M.D., Morais, C.T.D., Souza, M.G.D.: Padronização de metadados adotados no portal agregador de conteúdo científico oasisbr. Ciência da Informação 54(2) (2025). https://doi.org/10.18225/ci.inf.v54i2.7223
13. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. **31** (2018)
14. Rahman, M.M., Shiplu, A.I., Watanobe, Y.: Commentclass: a robust ensemble machine learning model for comment classification. Int. J. Comput. Intell. Syst. **17** (2024). https://doi.org/10.1007/s44196-024-00589-3
15. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002). https://doi.org/10.1145/505282.505283
16. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49 (1999). https://doi.org/10.1145/312624.312647

An Empirical Study of Compression Strategies for Tricorder: From General-Purpose Compressors to Shannon Entropy

Caio Guimaraes Herrera(✉), Hugo H. Nakamura, Isaac S. Soares, Simone R. S. Souza, and Paulo Sergio Lopes de Souza

University of Sao Paulo (ICMC/USP), Sao Carlos, SP, Brazil
caio.herrera@outlook.com

Abstract. Software Reliability Engineering (SRE) requires robust mechanisms to detect anomalies in production systems. The Tricorder methodology addresses this need by monitoring resource consumption and applying the DAMICORE clustering algorithm, which relies on the Normalized Compression Distance (NCD). However, computing the NCD is an expensive step, and its output is ultimately discarded, motivating the investigation of more efficient alternatives. This paper evaluates a set of compression strategies for the NCD metric in the context of Tricorder: seven general-purpose and multimedia compressors (ZLIB, GZIP, PPMD, BZ2, PNG, WEBP, JP2, HEIF), as well as an entropy-based approach (ENTROPY) that estimates compressed size from Shannon Entropy without producing compressed output. We conducted experiments using monitoring profiles from the Decision Tree application, with seven types of injected software faults evaluated against standard classification metrics (accuracy, precision, recall, and F1-score) and response time. Results show that ZLIB, GZIP, PPMD, PNG, and ENTROPY all achieve maximum effectiveness, correctly detecting all anomalies without false positives. ENTROPY additionally presents response times and per-iteration times consistently below the median across all anomaly types, matching or outperforming traditional compressors in efficiency. Most multimedia compressors introduced false positives in the control group, limiting their applicability. These findings establish Shannon Entropy as a viable, low-overhead alternative to traditional compression for NCD-based anomaly detection, particularly in resource-constrained or operationally restricted environments.

Keywords: Software reliability · Compression strategies · Anomaly detection

1 Introduction

Software Reliability Engineering (SRE) aims to ensure that production systems maintain high levels of availability, performance, and resilience. Several failures,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 243–260, 2027.
https://doi.org/10.1007/978-3-032-30494-0_16

however, do not manifest in conventional tests carried out in development environments, emerging only under real operational conditions [3].

The **Tricorder** methodology contributes to SRE in this context, as it enables the detection of anomalies during the execution of production systems [20]. **Tricorder** has two main phases: first, it monitors the software in production; then, it analyzes the collected consumption profiles through unsupervised machine learning. The latter is done by using **DAMICORE**, a methodology that relies on compressors, or compression strategies, to cluster monitoring profiles and determine a possible anomaly [23]. This step is of great importance, since the cost of compressors may hinder the widespread use of **Tricorder**.

In the **DAMICORE** methodology, compressors are used to calculate the Normalized Compression Distance (**NCD**) from the size of the compressed data. This is the most costly step in its execution. In contrast, the resulting compressed content is discarded as it has no utility in the anomaly detection process. This highlights an opportunity to reduce or eliminate the overhead of generating compressed output through alternative compression strategies.

This paper presents a comprehensive evaluation of alternative compressors for computing the **NCD** metric, as well as an entropy-based strategy for estimating the size of compressed data without generating the output. It compares these approaches with the ones used in previous studies, considering both effectiveness and efficiency in anomaly detection.

The adopted methodology consists of evaluating different compression strategies in the context of **Tricorder**, based on monitoring data previously collected from executions of an application with and without injected faults. The consumption profiles containing the monitoring data were clustered using **DAMICORE**, and each execution's results and response times were stored for analysis. The evaluation comprised both standard classification metrics (accuracy, precision, recall, and F1-score) and response time.

This paper makes two main contributions to unsupervised anomaly detection in the SRE domain. First, it provides a systematic comparison of nine compression strategies for NCD-based anomaly detection, from general-purpose and multimedia to entropy-based approaches; Second, it presents empirical evidence that *Shannon Entropy* is a viable alternative to traditional compressors, maintaining full detection accuracy while reducing computational overhead.

This paper is organized as follows. Section 2 describes the Tricorder methodology, used as the basis for anomaly detection. Section 3 presents the related work from the literature. Section 4 details the evaluated compression strategies, including traditional compressors and the alternative approach based on entropy calculation. Section 5 presents the experimental design, Sect. 6 shows the experimental results, and Sect. 7 analyzes and interprets the findings. Finally, Sect. 8 discusses the findings and limitations of this work, as well as directions for future research.

2 Background

Tricorder is a methodology for detecting software anomalies in production, based on the analysis of resource consumption profiles [20]. These profiles are represented by CSV files, created during the monitoring of an application's execution, containing resource usage metrics such as CPU and GPU utilization, RAM usage and I/O usage.

An initial version of the application, with no known faults, is considered the baseline for the methodology. From this version onward, resource consumption profiles are collected as a reference for anomaly detection. Figure 1 illustrates the behavior of **Tricorder** in this context.

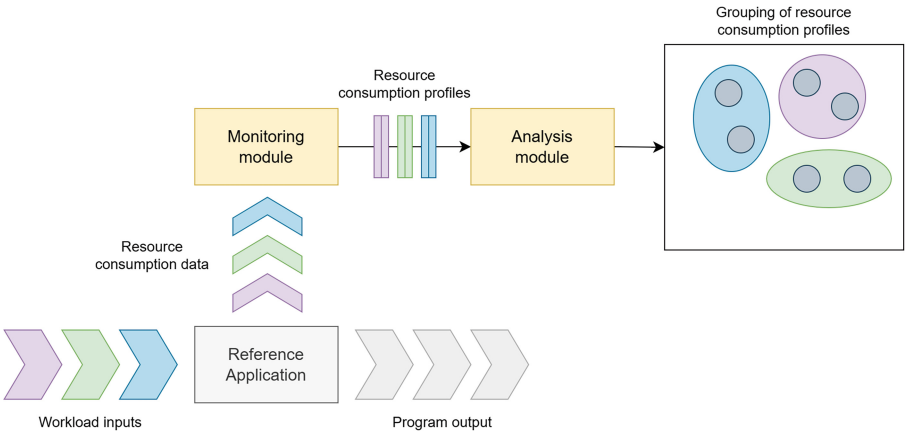


Fig. 1. Tricorder expected behavior when running a reference application. Adapted from Herrera et al. [13].

Then, profiles of a new version of the application, obtained after a maintenance cycle or a new release, are collected and grouped with their predecessors from the baseline version, generating clusters of application behaviors. The **Tricorder** methodology indicates a possible anomalous execution when a cluster made exclusively of profiles from the new version of the application is found. This process is illustrated in Fig. 2.

The clustering approach used by the **Tricorder** methodology is **DAMICORE** [23], a methodology to cluster generic data, through its *Python* implementation, *DamicorePy* [8]. This approach consists of three steps:

1. Construction of a distance matrix for each pair of inputs;
2. Transformation of the distance matrix into a phylogenetic tree;
3. Hierarchical clustering from the phylogenetic tree.

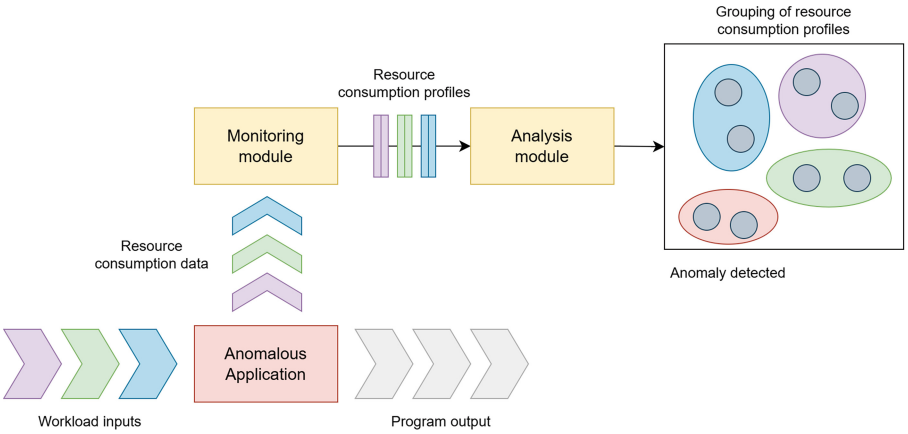


Fig. 2. Tricorder expected behavior when running an anomalous application. Adapted from Herrera et al. [13].

Constructing the distance matrix is the most time-consuming step. It is performed by calculating the **Normalized Compression Distance** (**NCD**) [9,18] between each pair of profiles. The **NCD** is a quantitative similarity metric, calculated between each pair of files, based on their compression. It operates by separately computing the compressed sizes of both files and their concatenation, as presented in Eq. 1.

$$NCD_z(x,y) = \frac{Z(xy) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}} \tag{1}$$

Here, x and y represent the input files, xy their concatenation, and $Z(\alpha)$ the function that returns the compressed size of a file α. The resulting value $NCD_z(x,y) \in [0,1]$ where 0 indicates maximum similarity and 1 indicates minimum similarity.

The performance of the **Tricorder** methodology is closely associated with the choice of compressor used to compute the **NCD** metric, since it determines the generation of the distance matrix and, consequently, the quality of the clusters, directly affecting the detection of software anomalies.

The **Tricorder** methodology, described here, is different from the Tricorder platform proposed by Sadowski et al. [21]. The latter is a supporting platform for developers when conducting program static analysis, and such a project is not related to the methodology we use; it shares only the name.

3 Related Work

The scientific literature establishes that selecting a data compressor is a complex task with no global optimal solution. Performance is dictated by a fundamental trade-off between compression ratio and response time for compression [14],

and it is highly dependent on the intrinsic characteristics of the data, such as repetitiveness and block size [29].

This complexity is also present in the use of compressors in similarity metrics, such as the **NCD**, which is central to this study. Theory suggests that properties of compressors, such as idempotency ($C(xx) \approx C(x)$), are crucial for the robustness of **NCD**. However, practical limitations, such as the compression window size, may introduce significant deviations [7]. Indeed, empirical investigations demonstrate that good theoretical performance of a compressor does not guarantee the best results in practice, underlining the need for contextualized validation [1].

The theoretical pillar of **NCD** is its role as a practical approximation of *Kolmogorov Complexity*, which describes the information contained in an object by the size of the smallest program that generates it. The fundamental relationship between this theory and *Shannon Entropy* is explored in depth by Teixeira et al. [27]. Their work reinforces the theorem that establishes an equivalence between the expected value of *Kolmogorov Complexity* and *Shannon Entropy* for a computable distribution, formalized by the relation presented in Eq. 2. The main conclusion of their study is that this elegant equivalence is an exclusive property of *Shannon Entropy*, not holding for other generalizations such as Rényi and Tsallis entropies, which provides theoretical support for the approach proposed in this paper.

$$0 \leq \sum_{x} P(x)K(x) - H(P) \leq K(P) \tag{2}$$

These challenges have a direct practical impact on anomaly detection. The works of Braga [5] and Durães [11], for example, apply the **Tricorder** methodology with different compressors, highlighting how the choice of the compression algorithm directly impacts performance. The analysis of different compressors for applications such as the *Decision Tree* [24], used in Braga's experiments, provided the direct motivation for the in-depth comparative investigation carried out in this paper.

In this context, the investigations presented here differ from related work by conducting an in-depth empirical analysis. This paper investigates the impact of a set of compressors, lossless and lossy, as well as a new approach, which employs the calculation of *Shannon Entropy* as a replacement for traditional compressors, in computing the **NCD**. Additionally, the effectiveness and efficiency of the **Tricorder** methodology are evaluated to fill a gap in its validation and to expand the understanding of the impact of different approaches on anomaly detection.

4 Compression Strategies

For the experimental stage of this work, a set of compressors was defined, based on a selection that prioritized two fundamental criteria: the representativeness of the main algorithmic paradigms (dictionary, block transformation, and statistical [22]) and the relevance of each compressor as a *de facto* standard in

computational practice [14]. Both lossless and lossy compressors were chosen to investigate whether lossy strategies could also be viable for **Tricorder** and to determine the benefits of their use in terms of effectiveness and efficiency.

Within the dictionary-based compression paradigm, the *DEFLATE* algorithm [17] operates by combining the *LZ77* technique [30] with *Huffman coding* [15]. *LZ77* replaces repeated data sequences with references to previous occurrences, using a sliding window to keep track of recently processed data. When it encounters a previously seen sequence, the algorithm replaces it with a pair of values indicating the distance and length of the original sequence [30]. Then, it compresses the remaining symbols, the literal characters, and the references generated by *LZ77* using *Huffman coding*, assigning variable-length codes based on frequency of occurrence [15]. Its main implementations, the **ZLIB** library and the **GZIP** tool [10], consolidated this approach as a *de facto* standard, being fundamental in countless contexts, from web data transfer to the structuring of different file formats [12]. The **PNG** format [4], a lossless compression format also based on the *DEFLATE* algorithm, applies predictive filters as preprocessing that transform image data to highlight repetitive patterns, thereby improving its compression efficiency.

The **WEBP** compressor supports two modes: lossy, based on the intra-frame prediction of the *VP8* codec [2], and lossless, which uses spatial prediction, color cache, and blocks. In lossless mode, these techniques reduce image-specific redundancies before the final compression with *DEFLATE*, similar to **ZLIB**. Spatial prediction estimates pixels from their neighbors, the cache reuses recent colors, and blocks capture local patterns, achieving higher compression than simple repeated-sequence substitution.

In the block transformation paradigm, Bzip2 (**BZ2**) [25] serves as a reference for its high compression ratios [14]. Its approach consists of applying the Burrows-Wheeler Transform (BWT) [6], a reversible permutation that reorders the characters of a block based on the lexicographic ordering of all its cyclic rotations. This process, although it does not compress the data directly, effectively groups identical symbols, creating high locality of reference. The transformed block is then processed by the Move-to-Front (MTF) transformation [6], which generates a sequence of low-value integers, efficiently compressed at the end using *Huffman coding*.

As a representative of the statistical paradigm, the Prediction by Partial Matching algorithm (**PPM**) [19] operates predictively, building a dynamic probabilistic model from the data stream. To encode a symbol, **PPM** analyzes the immediately preceding characters (the "context") to calculate the probability of the next symbol, which is then encoded with fewer bits by an arithmetic coder. One of its best-known implementations is **PPMd**. The effectiveness of **PPM** on data with clear statistical structure, such as text, makes it a key reference for evaluating contextual information modeling [22].

Beyond the paradigms described above, several image compressors rely on transform-based approaches. The **JPEG 2000** (**JP2**) compressor is an evolution of **JPEG**, offering both lossy and lossless compression, support for larger

images, progressive decoding, greater error tolerance, and processing efficiency. It uses the *Discrete Wavelet Transform* (DWT) for lossy compression and the *Integer Reversible Wavelet Transform* for lossless compression, allowing it to switch between maximum quality and high compaction. After the transformation, the coefficients are quantized (in lossy mode) and encoded by *EBCOT* (Embedded Block Coding with Optimal Truncation), which organizes independent blocks and generates a scalable stream [16].

The **HEIF** (High Efficiency Image File Format) uses the *HEVC/H.265* codec for lossy compression, segmenting images into blocks, applying transforms to remove redundancies, and encoding coefficients by entropy [28]. Unlike **ZLIB**, a generic lossless compressor, **HEIF** is optimized for images, achieving much higher compression ratios with controlled loss while also supporting multiple images, transparency, and metadata.

In contrast to compressors that estimate complexity through algorithmic heuristics, this work evaluates another approach, here referred to as **ENTROPY**, which represents a purely statistical measure of complexity. Grounded in Claude Shannon's seminal work on Information Theory [26], this measure corresponds to the calculation of zero-order *Shannon Entropy*. Entropy, $H(X)$, quantifies the uncertainty or "information content" of a data source based solely on the probability distribution $p(x_i)$ of its symbols, as expressed in Eq. 3.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{3}$$

Unlike compressors, which search for specific patterns (such as literal repetitions or contexts), **ENTROPY** provides a mathematical measure that summarizes the information content of the file, without producing compressed output. Table 1 summarizes the evaluated compressors discussed in this section.

Table 1. Selected compression strategies to evaluate against the **NCD** metric.

Name	Used for	Paradigm	Data loss
ZLIB	Generic data	Dictionary-based	Lossless
GZIP	Generic data	Dictionary-based	Lossless
PPMD	Generic data	Statistical	Lossless
BZ2	Generic data	Block transformation	Lossless
PNG	Multimedia	Dictionary-based	Lossless
WEBP-LS	Multimedia	Dictionary-based	Lossless
WEBP-LY	Multimedia	Block transformation	Lossy
JP2-LS	Multimedia	Block transformation	Lossless
JP2-LY	Multimedia	Block transformation	Lossy
HEIF	Multimedia	Block transformation	Lossy
ENTROPY	Theoretical approximation	Statistical measure	N/A

5 Experimental Studies

We conducted experiments inside an Ubuntu 22.04.5 LTS environment running via WSL2 (Windows Subsystem for Linux 2). It was hosted on a computer with an 8-core Intel Core i7-10510U @ 1.80 GHz, 8 GB DDR4 RAM, and Windows 11 (24H2).

The experiment conducted in this paper used monitoring data already made available by Braga [5]. The profiles refer to the *Decision Tree* application, loaded with the *Breast Cancer* and *Phoneme* datasets [24]. The *Decision Tree* application was chosen because it has heavy workloads (longer response times) and presents stochastic behavior, which makes the clustering performed by Tricorder non-trivial [5]. From the original *Decision Tree* code, different faulty versions were created, each with the insertion of a specific software fault:

- **LOGIC**: logic errors. Low impact on resource consumption;
- **MODEL**: machine learning configuration problems. Low impact on resource consumption;
- **MEMORY**: memory usage problems. Low impact on resource consumption;
- **API**: missing or misplaced API calls. Medium impact on resource consumption;
- **TRAIN**: missing training data. Medium impact on resource consumption;
- **CONC**: synchronization problems. High impact on resource consumption; and
- **PROCESS**: problems in object initialization or method usage. High impact on resource consumption.

For clustering to be performed appropriately, it was necessary to obtain consumption profiles that reflect the expected execution of the selected applications. Therefore, the **CONTROL** and **ANOTHER-CONTROL** types were defined, where **CONTROL** is used as a baseline, and **ANOTHER-CONTROL** is used to verify false positives in anomaly detection. The expected result is that **ANOTHER-CONTROL** is not flagged by **Tricorder** as an anomaly.

In the context of **Tricorder**, the anomaly detection process is executed incrementally. Each profile is inserted one at a time, and **DAMICORE** performs clustering after each insertion. This process continues until a group consisting exclusively of consumption profiles from an anomaly type is identified, or until all available profiles have been included. Thus, the experiment consists of the following steps:

1. Selection of consumption profiles of a given anomaly type;
2. Generation of reference groups with control consumption profiles;
3. For each new anomalous consumption profile loaded, clustering is executed again to update the groups;
4. If an exclusive group is found, i.e., an anomaly detected, the process is interrupted, and the current iteration is saved, along with the system's total response time;

5. Steps 1–4 are repeated until all profiles of each type have been used, and the full experiment is executed 30 times per anomaly type.

The evaluation of effectiveness regarding the selected compression strategies considered four main metrics:

1. **Accuracy:**
$$\frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$
2. **Precision:**
$$\frac{TP}{TP + FP} \tag{5}$$
3. **Recall:**
$$\frac{TP}{TP + FN} \tag{6}$$
4. **F1-Score:**
$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Table 2 presents the definition of the confusion matrix for the effectiveness metrics. The evaluation of efficiency regarding compression strategies measures the total response time of the detection process, independently of whether the detection was successful. Furthermore, to evaluate the set of compressors in terms of effectiveness, we considered the number of necessary iterations to detect an anomaly.

Table 2. Confusion matrix for the anomaly detection task performed by **Tricorder**.

	Predicted Anomalous	Predicted Normal
Actual Anomalous	True Positive (TP)	False Negative (FN)
Actual Normal	False Positive (FP)	True Negative (TN)

To properly use the image compressors, however, the input data must be transformed, since the original monitoring files are binary and not suitable for these compressors. The files were then transformed into non-compressed BITMAP files, where each set of 3 bytes represents a pixel's color: the first byte is the red component, the second is the green component, and the third is the blue component. It is worth noting that the chosen encoding is arbitrary with respect to image structure: the resulting bitmaps carry no spatial semantics, which may affect the behavior of compressors designed to exploit two-dimensional visual redundancies.

6 Results

The experimental results are presented below, organized into graphs and tables to highlight the performance of the compression strategies and the differences among them. Table 3 shows the median number of iterations required to detect each anomaly type per compression strategy. The control group is also shown, identified as **A.C.** (from **ANOTHER-CONTROL**) due to space limitations. Cells highlighted in green have a number of iterations that is less than or equal to the median value for the anomaly type across all compression strategies. The median for the control group was not considered, as any detection on the **A.C.** represents a false positive.

Table 3. Median number of iterations for each compressor to signal the detection of an anomaly.

Compressor	A.C.	MEMORY	MODEL	LOGIC	TRAIN	API	CONC	PROCESS
ZLIB	-	27	24	8	17	16	4	4
PPMD	-	17	24	16	6	4	17	2
GZIP	-	29	20	7	17	16	4	2
BZ2	-	27	17	7	9	-	10	2
PNG	-	18	10	9	8	8	5	2
WEBP-LS	16	-	-	9	20	25	11	7
WEBP-LY	19	18	13	25	4	9	3	4
JP2-LS	12	14	5	16	13	7	5	7
JP2-LY	12	14	5	16	13	7	5	7
HEIF	11	2	6	7	4	5	15	1
ENTROPY	-	15	7	24	6	24	11	4
Median	-	**17.5**	**11.5**	**9**	**9**	**8.5**	**5**	**4**

Almost all compression strategies detected all anomalies, with the exception of **BZ2**, which did not detect the **API** anomaly, and **WEBP-LS**, which failed to detect both **MEMORY** and **MODEL** anomalies. Additionally, compressors **WEBP** (both lossy and lossless), **JP2** (both lossy and lossless), and **HEIF** detected anomalies in the control group, resulting in false positives.

Regarding the iterations required to detect an anomaly, the **HEIF** compressor required the fewest iterations for anomalies **MEMORY**, **MODEL**, **LOGIC**, **TRAIN** and **PROCESS**. For **API**, the **PPMD** compressor took the minimum amount of iterations; for **CONC**, it was **WEBP-LY**.

Figure 3 presents the bar charts of the previously defined effectiveness metrics for each compression strategy applied to the anomaly detection task. We observe that the compression strategies **ZLIB**, **PPMD**, **GZIP**, **PNG**, and **ENTROPY** achieved the best results across all metrics, indicating they correctly detected the anomalous profiles and distinguished them from the control profiles. Image compressors, with the exception of **PNG**, are shown to be affected by false positives, leading to reduced accuracy, precision, and F1-score.

WEBP-LS had the lowest results across all metrics due to a combination of detecting a false positive and missing two true positives.

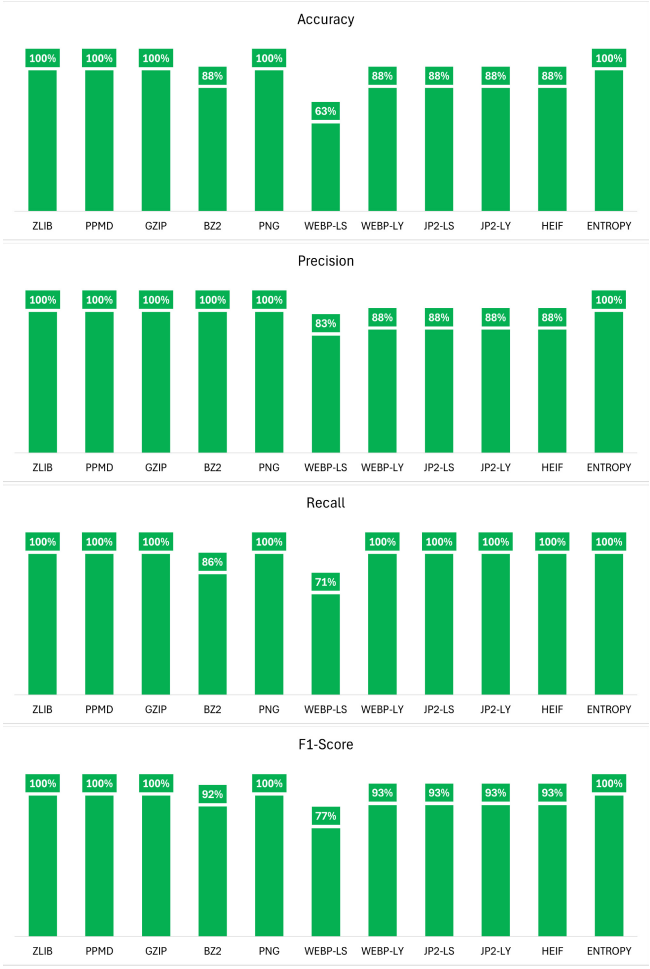


Fig. 3. Bar plots of the effectiveness metrics for each compression strategy in the Decision Tree application.

The efficiency counterpart of this experiment is shown in Table 4, which highlights the median response time for each compression strategy to detect an anomaly. The results shown are only for the used anomalies, and do not include the control group. This metric considers the total execution time for *Damicorepy*, including the time required to compress and group the profiles. For the cases where an anomaly was not detected, the table presents the total execution time. In the same way, cells in which the response time was less than or equal to the median response time for that anomaly are highlighted in green.

Table 4. Median response time (seconds) for each compressor per anomaly type.

Compressor	MEMORY	MODEL	LOGIC	TRAIN	API	CONC	PROCESS
ZLIB	4.28	3.67	1.06	2.33	2.23	0.60	0.58
PPMD	3.47	5.59	3.27	1.13	0.83	3.48	0.52
GZIP	11.88	6.71	2.05	5.45	5.13	1.27	0.78
BZ2	6.22	3.26	1.31	1.61	7.08	1.85	0.50
PNG	14.24	6.82	6.13	5.24	5.45	3.44	1.76
WEBP-LS	38.42	38.30	8.37	21.22	31.02	10.71	6.21
WEBP-LY	18.46	12.14	29.42	3.81	8.16	3.18	3.74
JP2-LS	16.53	5.70	19.46	14.39	7.68	6.15	7.19
JP2-LY	16.30	5.54	19.23	14.16	7.58	5.65	7.15
HEIF	12.66	28.64	33.60	19.79	24.63	82.60	9.01
ENTROPY	2.67	1.21	4.91	1.03	4.91	1.90	0.74
Median	**12.66**	**5.70**	**6.13**	**5.24**	**7.08**	**3.44**	**1.76**

We can observe that compressors **ZLIB**, **BZ2** and **ENTROPY** had faster response times than the median of all compressors for every type of anomaly. **PPMD**, **GZIP** and **PNG** also had response times faster than the median for the majority of the anomalies. Overall, **ENTROPY** was the fastest compression strategy for anomalies **MEMORY**, **MODEL** and **TRAIN**. Compressor **ZLIB** was the fastest for **LOGIC** and **CONC**. **PPMD** had the shortest response time for **API** and **BZ2** was the fastest for the **PROCESS** anomaly.

Going further into this analysis, Table 5 accounts for the time spent per iteration, including a single round of compression and grouping. In this analysis, we can see a more consistent behavior across compression strategies, where the **PNG** compressor is always at the median time, where compressors **ZLIB**, **PPMD**, **GZIP**, **BZ2** and **ENTROPY** are faster than the median iteration time for all evaluated anomalies. **HEIF** is the compressor with the highest iteration time for all scenarios, despite requiring fewer iterations to detect the majority of anomalies.

From these tables, we can begin to understand the differences in response time across compressors. However, it is essential to perform a statistical analysis to detect significant differences in the data. For this reason, the Shapiro-Wilk test was used to verify the normality of the data. Since the majority of pairs of anomaly type and compressor follow nonparametric distributions, the nonparametric Wilcoxon test was applied to confirm significant differences in response times across compression strategies.

All comparisons with $p < 0.05$ were considered statistically significant. Table 6 presents the compression strategy pairs that resulted in non-significant differences in response time according to the Wilcoxon test, after normalization through *Holm's method* to reduce the false discovery rate due to the number of conducted experiments. All other compressor combinations, which accounted for the majority of the analysis, showed significant differences.

Table 5. Median iteration time (seconds) for each compressor per anomaly type.

Compressor	MEMORY	MODEL	LOGIC	TRAIN	API	CONC	PROCESS
ZLIB	0.16	0.15	0.13	0.14	0.14	0.15	0.15
PPMD	0.20	0.23	0.20	0.19	0.21	0.20	0.26
GZIP	0.41	0.34	0.29	0.32	0.32	0.32	0.39
BZ2	0.23	0.19	0.19	0.18	0.24	0.18	0.25
PNG	0.79	0.68	0.68	0.66	0.68	0.69	0.88
WEBP-LS	1.28	1.28	0.93	1.06	1.24	0.97	0.89
WEBP-LY	1.03	0.93	1.18	0.95	0.91	1.06	0.93
JP2-LS	1.18	1.14	1.22	1.11	1.10	1.23	1.03
JP2-LY	1.16	1.11	1.20	1.09	1.08	1.13	1.02
HEIF	6.33	4.77	4.80	4.95	4.93	5.51	9.01
ENTROPY	0.18	0.17	0.20	0.17	0.20	0.17	0.19
Median	**0.79**	**0.68**	**0.68**	**0.66**	**0.68**	**0.69**	**0.88**

Table 6 should be analyzed in conjunction with the time distributions presented in Fig. 4, that provides a visual representation of the response time data for each compression strategy, when looking for three representative anomaly types: **API**, **MEMORY** and **PROCESS**. It visually confirms the statistical similarities found in the Wilcoxon test results.

Table 6. Results for non-significant differences in response time among compression strategies.

Comparison	Anomaly	P-Value
BZ2/ENTROPY	ANOTHER-CONTROL	0.314166
BZ2/JP2-LS	API	1.000000
BZ2/JP2-LY	API	1.000000
BZ2/PPMD	ANOTHER-CONTROL	0.113038
GZIP/HEIF	MEMORY	0.464772
GZIP/PNG	API	0.538887
GZIP/PNG	MODEL	0.334234
GZIP/PNG	TRAIN	0.396577
JP2-LS/JP2-LY	ANOTHER-CONTROL	1.000000
JP2-LS/JP2-LY	CONC	0.132791
JP2-LS/JP2-LY	LOGIC	0.526366
JP2-LS/JP2-LY	PROCESS	1.000000
JP2-LS/WEBP-LY	MEMORY	0.497101
JP2-LS/PPMD	MODEL	1.000000
JP2-LY/PPMD	MODEL	0.243841
PNG/PPMD	CONC	1.000000

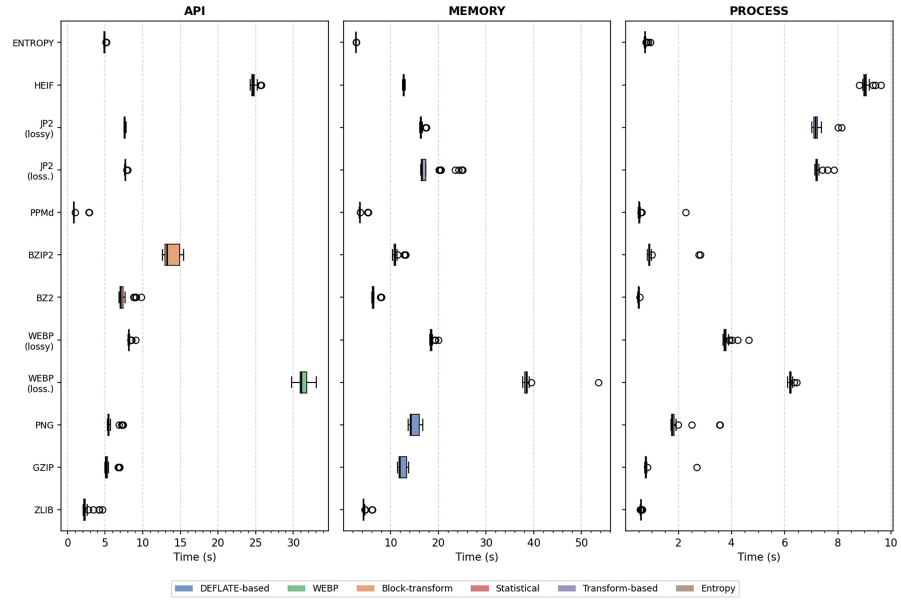


Fig. 4. Boxplots for response time of each compressor against three types of anomalies.

7 Discussion

Results shown in Sect. 6 reveal a consistent correlation between the impact level of a fault and the number of iterations required to detect it. Low-impact anomalies (**MEMORY**, **MODEL**, and **LOGIC**) required an average of 14.7 iterations to be detected, medium-impact anomalies (**TRAIN** and **API**) an average of 11.3, and high-impact anomalies (**CONC** and **PROCESS**) only 6, as the larger changes they introduce in resource consumption profiles make clusters more distinct earlier.

These results reiterate those from previous works regarding the efficiency and effectiveness of both **ZLIB** and **PPMD** compressors, as discussed by Braga [5] and Durães [11]. Moreover, the **PNG** compressor showed positive results: both in effectiveness, by detecting all anomalies without false positives, and in efficiency, with response times below the median for five out of seven anomalies. When looking at iteration time, **PNG** was exactly at the median time for all anomalies.

Some of the non-significant pairs have algorithmic explanations. The similarity between **JP2-LS** and **JP2-LY** is natural, as they are two modes of the same codec and their execution cost difference is negligible. Similarly, **GZIP** and **PNG** both rely on *DEFLATE*, so comparable performance is expected.

WEBP, **JP2** and **HEIF** detected false positives for the control group, while **WEBP-LS** failed to detect two anomalies. A likely contributing factor is the BITMAP transformation: image-specific codecs exploit spatial redundancies in

visual data, and when applied to arbitrarily encoded monitoring bytes, they may perceive distinct consumption profiles as structurally similar, causing the NCD to underestimate distances and triggering spurious detections. This suggests that the preprocessing step is a key limitation for these compressors in the **Tricorder** context, and both the encoding scheme and alternative representations should be explored in future work. Despite that, the **HEIF** compressor required the minimum amount of iterations for four out of seven anomalies, and had the number of iterations below the median in two other anomalies, which indicates that fewer profiles are required for it to detect an anomaly, even if this detection takes a longer time.

The **ENTROPY** compression strategy yielded positive results by detecting all anomalies without generating false positives, while also maintaining total response time and iteration time below the median for all anomalies. This combination of results is matched only by the **ZLIB** compressor, which supports the use of *Shannon Entropy* as an alternative to traditional compression strategies. Furthermore, although it requires more iterations to detect **LOGIC**, **API**, and **CONC** anomalies, its short iteration times keep the total response time below the median.

8 Conclusion and Future Work

This paper evaluated alternative compression strategies for computing the **NCD** metric in the context of the **Tricorder** methodology, including an entropy-based approach (**ENTROPY**) that estimates compressed size without generating actual compressed output. Seven general-purpose and multimedia compressors (**ZLIB**, **GZIP**, **PPMD**, **BZ2**, **PNG**, **WEBP**, **JP2**, and **HEIF**) were compared against **ENTROPY** with respect to both effectiveness and efficiency in anomaly detection. The experiments used monitoring profiles from the *Decision Tree* application, covering seven types of injected software faults. They evaluated results using standard classification metrics (accuracy, precision, recall, and F1-score) and response-time analysis.

Results indicated that the **ENTROPY** compression strategy is efficient and effective in detecting anomalies, comparable to other compression approaches explored in the paper. It correctly identified all anomalies and performed similarly to previously adopted strategies, such as **ZLIB** and **PPMD**.

Regarding efficiency, the **ENTROPY** strategy consistently achieved among the lowest iteration times across all anomaly types, ranking second only to **ZLIB** for most scenarios. This characteristic makes it a promising alternative, especially for detecting low-impact anomalies, as the number of iterations required to detect an anomaly is directly related to the anomaly's impact on resource consumption. Thus, low-impact anomalies benefit more from strategies with low iteration times.

In addition to its effectiveness and efficiency results, the simplicity of implementation and independence from external libraries make **ENTROPY** a low-cost and promising alternative for resource-constrained or operationally restricted environments.

From the perspective of image compressors, **PNG** performed well, detecting all anomalies without false positives. Moreover, it showed response times below the median for five of seven anomalies, while its iteration time was equal to the median for all seven anomalies. These results are not as good as those of established compressors, such as **ZLIB** and **PPMD**, but they are a viable alternative.

Although less precise and with higher execution time, **HEIF** showed an advantage by requiring fewer iterations, making it promising in applications where the number of operations is more critical than total time. These findings reinforce that the choice of compressor should be based not only on precision and execution time, but also on the constraints and priorities of each use scenario. Ultimately, the choice of strategy depends on the specific constraints and priorities of the target application.

The consumption profiles used in the experiment were obtained by monitoring a single application with a subset of possible anomalies. The results may not fully reflect the typical behavior of the evaluated compression strategies across other subsets of anomalies, given the variability in the results presented. Future work may strengthen the results of this study by diversifying the application domains to verify the generalization of compression strategies and reduce biases arising from analyzing a single application. It is also important to conduct robust experiments focused on verifying false-positive rates, with a greater number of alternative control profiles.

Acknowledgments. The authors would like to thank CAPES (PROEX), CNPq (306719/2025-8, 311207/2025-1, 406941/2025-4), and FAPESP (2013/07375-0, 2019/06937-0, 2019/26702-8) for their financial support.

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

References

1. Axelsson, S., Bajwa, K.A., Srikanth, M.V.: File fragment analysis using normalized compression distance. In: Peterson, G., Shenoi, S. (eds.) DigitalForensics 2013. IAICT, vol. 410, pp. 171–182. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41148-9_12
2. Bankoski, J., Koleszar, J., Quillio, L., Salonen, J., Wilkins, P., Xu, Y.: VP8 Data Format and Decoding Guide. Request for Comments RFC 6386, Internet Engineering Task Force (IETF) (2011). https://www.rfc-editor.org/rfc/rfc6386
3. Beyer, B., Jones, C., Petoff, J., Murphy, N.: Site Reliability Engineering: How Google Runs Production Systems. O'Reilly (2016)
4. Boutell, T., Lane, T., et al.: PNG (Portable Network Graphics) Specification 1.0. Request for Comments RFC 2083, Internet Engineering Task Force (IETF) (1997)
5. Braga, D.: Avaliação do uso de agrupamento de dados de desempenho para apoiar o teste de software no domínio de aprendizagem de máquina. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos (2022). (in Portuguese). https://doi.org/10.11606/D.55.2022.tde-09112022-160513

6. Burrows, M., L, D.J.W.D.I.G.I.T.A., Taylor, R.W., Wheeler, D.J., Wheeler, D.: A block-sorting lossless data compression algorithm (1994). https://api.semanticscholar.org/CorpusID:2167441
7. Cebrián, M., Alfonseca, M., Ortega, A.: Common pitfalls using the normalized compression distance: what to watch out for in a compressor. Commun. Inf. Syst. **5**(4), 367–384 (2005). https://doi.org/10.4310/CIS.2005.v5.n4.a1
8. Cesar, B.K.M.: Estudo e extensão da metodologia DAMICORE para tarefas de classificação. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil (2016). https://doi.org/10.11606/D.55.2016.tde-16112016-170837
9. Cilibrasi, R., Vitányi, P.M.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005). https://doi.org/10.1109/TIT.2005.844059
10. Deutsch, P.: GZIP file format specification version 4.3. Request for Comments RFC 1952, Internet Engineering Task Force (IETF) (1996). https://www.rfc-editor.org/rfc/rfc1952
11. Durães, T.d.J.O.: Avaliação do uso de perfis de desempenho aplicados no teste e software em diferentes domínios de aplicação. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos (2022). (in Portuguese). https://doi.org/10.11606/D.55.2022.tde-05042023-144747
12. Gailly, J.l., Adler, M.: Zlib applications (2026). https://zlib.net/apps.html, Accessed 20 Mar 2026
13. Guimaraes Herrera, C., Silva Montes, V., Souza, S.R.S., Ribeiro dos Santos, R., Lopes de Souza, P.S.: An architecture for a reliability tool applied to distributed systems. In: Computational Science and Its Applications – ICCSA 2025, pp. 332–348 (2025)
14. Gupta, A., Bansal, A., Khanduja, V.: Modern lossless compression techniques: review, comparison and analysis. In: 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–8 (2017). https://doi.org/10.1109/ICECCT.2017.8117850
15. Huffman, D.A.: A method for the construction of minimum-redundancy codes. Proc. IRE **40**(9), 1098–1101 (1952)
16. ISO/IEC-JTC-1/SC-29: Information technology – jpeg 2000 image coding system: core coding system. Standard ISO/IEC 15444-1:2004, Int. Org. for Standardization (2004)
17. Katz, P.: DEFLATE Compressed Data Format Specification version 1.3. Request for Comments RFC 1951, Internet Engineering Task Force (IETF) (1996). https://www.rfc-editor.org/rfc/rfc1951
18. Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004). https://doi.org/10.1109/TIT.2004.838101
19. Moffat, A.: Implementing the PPM data compression scheme. IEEE Trans. Commun. **38**(11), 1917–1921 (1990)
20. Montes, V.S., Herrera, C.G., Durães, T.J.O., Delbem, A.C.B., de Souza, S.R.S.L., Souza, P.S.L.d.: Software reliability forecasting using anomalies in profiles of resource usage. In: 2025 IEEE 36th International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 189–196 (2025). https://doi.org/10.1109/ISSREW67781.2025.00071
21. Sadowski, C., Van Gogh, J., Jaspan, C., Soderberg, E., Winter, C.: Tricorder: building a program analysis ecosystem. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, pp. 598–608 (2015). https://doi.org/10.1109/ICSE.2015.76

22. Sakr, S.: Xml compression techniques: a survey and comparison. J. Comput. Syst. Sci. **75**(5), 303–322 (2009). https://doi.org/10.1016/j.jcss.2009.01.004, https://www.sciencedirect.com/science/article/pii/S0022000009000142
23. Sanches, A., Cardoso, J.M., Delbem, A.C.: Identifying merge-beneficial software kernels for hardware implementation. In: 2011 International Conference on Reconfigurable Computing and FPGAs, pp. 74–79 (2011). https://doi.org/10.1109/ReConFig.2011.51
24. Santos, S.A., Silveira, B., Durelli, V., Durelli, R., Souza, S., Delamaro, M.: On using decision tree coverage criteria for testing machine learning models. In: Proceedings of the 6th Brazilian Symposium on Systematic and Automated Software Testing. SAST '21, pp. 1–9. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3482909.3482911
25. Seward, J.: bzip2 and libbzip2, version 1.0.8. sourceware.org (2019). https://sourceware.org/bzip2/manual/manual.html. Accessed 20 Mar 2026
26. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
27. Teixeira, A., Matos, A., Souto, A., Antunes, L.: Entropy measures vs. Kolmogorov complexity. Entropy **13**(3), 595–611 (2011). https://doi.org/10.3390/e13030595
28. Telecommunication Standardization Sector of ITU: H.265 : High efficiency video coding. Recommendation ISO/IEC 23008-2:2013, ITU-T (2013)
29. Yao, K., Li, H., Shang, W., Hassan, A.E.: A study of the performance of general compressors on log files. Empir. Softw. Eng. **25**(4), 3043–3085 (2020). https://doi.org/10.1007/s10664-020-09822-x
30. Ziv, J., Lempel, A.: A universal algorithm for sequential data compression. IEEE Trans. Inf. Theory **23**(3), 337–343 (1977)

A New Training Method for Text Classification in Mental Health: LatentGLoss

Korhan Sevinç(✉)

TOBB University of Economics and Technology, Ankara, Turkey
ksevinc@etu.edu.tr

Abstract. Mental health text classification remains difficult when decisions must rely on short, naturally phrased user language. We study a seven-way task on a benchmark built by merging nine public mental-health text collections into one label schema. The paper reports a staged comparison: linear and ensemble models on TF–IDF features, convolutional networks, recurrent networks (including LSTM and GRU), and fine-tuned BERT-family encoders, so that performance can be read as a progression across representation choices. The main contribution is DualLatentGNet, a teacher–student framework in which a dual-objective CNN teacher (classification plus reconstruction) supervises a student. Unlike typical distillation pipelines that emphasize softened class targets, we route supervision through both the teacher's predictions and its latent geometry: a Gaussian mixture model is fit to concatenated teacher latent vectors and logits, and the student objective adds a latentG term tied to the assigned mixture component and to Euclidean distance in that feature space. On the merged benchmark, DualLatentGNet reaches 95.09% accuracy compared with 94.83% for the strongest fine-tuned BERT baseline under the same protocol, which supports the view that explicit latent alignment complements the baseline hierarchy we evaluate.

Keywords: mental health · text classification · teacher-student · knowledge distillation · latent representations

1 Introduction

Mental well-being promotion is central to the action plans of both the World Health Organization (WHO) for the years 2013–2020 [22] and the European Union's European Pact on Mental Health and Well-being [5]. The key breakthrough in addressing mental health issues lies in developing tools for early detection and preventive measures [8]. The WHO action plan highlights the importance of health policies and programs that not only address the needs of those affected by mental health disorders but also work to preserve mental well-being. These policies emphasize evidence-based, non-pharmacological interventions, focusing on early intervention and preventing unnecessary institutionalization and medicalization. Successful interventions are particularly dependent

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 261–273, 2027.
https://doi.org/10.1007/978-3-032-30494-0_17

on how frequently the therapy can be accessed [15]. In this context, automated systems have significant advantages over traditional therapies due to their ability to provide continuous support at minimal additional cost. Consequently, health assistants capable of delivering therapeutic interventions have gained considerable attention in recent years [1,7]. However, these systems are predominantly based on manually designed rules. In contrast, research in statistical methods for conversational systems has largely been limited to narrow-domain information-seeking dialogues [6,16,20,23].

We focus on inferring mental-health category cues from language people already produce—conversations, chat logs, forum posts, and similar text—because such material can surface signals between infrequent clinical contacts. For this work we did not collect new interviews; instead we merged nine publicly released datasets into one benchmark with a shared text field and seven labels, applied a single normalization and back-translation pipeline, and then trained a broad set of models so that conclusions are tied to one harmonized resource (Sect. 3). Classical machine learning on TF–IDF features, Word2Vec-based convolutional and recurrent models (including LSTM and GRU), and fine-tuned transformers (BERT and related encoders) are all included, which lets us relate our proposal to both lightweight and contextual baselines (Sect. 5).

We then describe DualLatentGNet: a dual-objective teacher network supplies class logits and internal representations, and a student is trained with a loss that combines the usual classification term with mixture-based and distance-based signals derived from the teacher's latent space (Sect. 4). The aim is to transfer information through predictions *and* through where those predictions live in representation space, rather than through logits-only distillation alone. Code, data links, training configurations, and model checkpoints are maintained at https://github.com/korhansevinc/LatentG-Loss.

2 Related Work

Interest in computational support for mental health has motivated many text-classification studies in which online writing is treated as a weak but abundant signal [2]. Cohan et al. analyzed mental-health forum posts with traditional machine-learning models, including support vector machines, and showed that lexical choices, sentiment, and simple frequency cues remain informative even though unstructured social language is noisy. Their discussion underscores how feature engineering and classifier choice interact when the target condition is expressed indirectly.

Recurrent architectures address limits of bag-of-words views by modeling token order. Rea et al. used LSTM networks for mental-health-related text and reported gains over classical baselines when sequential context is preserved [21]. Dey and Desai combined LSTMs with GloVe embeddings to obtain stronger semantic vectors when large transformer fine-tunes are expensive to deploy [9]. The CNN, RNN, LSTM, and GRU baselines we train in Sect. 5 sit in the same methodological line: they test how far convolutional and recurrent encoders can go without contextual pre-training.

Pre-trained transformers fine-tuned in domain have become the default strong comparators in NLP. Saleela et al. adapted BERT to mental-health text and reported strong discrimination across conditions such as depression, anxiety, and stress [17]. We mirror that practice by fine-tuning BERT, DistilBERT, ALBERT, and RoBERTa on our merged corpus (Table 1).

Multimodal systems widen the input beyond text: Kumar et al. combined written language with speech features to exploit acoustic correlates of distress alongside wording [4]. Our scope is deliberately text-only so that any gain from DualLatentGNet is attributable to the training objective and architecture rather than to additional sensors.

Across these lines of work, two constraints recur: labels and domains differ across corpora, and high accuracy from raw sentences remains challenging when classes are rare or language is indirect. Some pipelines therefore augment text with engineered linguistic or clinical features. Here we keep the input as plain text and ask whether a student model can absorb richer supervision from a teacher's internal representations as well as from its class predictions.

Teacher–student and distillation methods are familiar in vision and general NLP, but mental-health text classification has seen less work in which the student is explicitly regularized toward the *geometry* of the teacher's latent space via a mixture density over teacher features. DualLatentGNet (Sect. 4) targets that gap: a Gaussian mixture is fit to concatenated teacher latents and logits, and density- and distance-based terms enter the student loss alongside standard classification, which also offers an additional handle when training data are finite relative to model capacity.

3 Dataset

The dataset for this study combines nine publicly available mental-health-oriented text resources: *Depression Reddit Cleaned*, *Human Stress Prediction*, *Predicting Anxiety in Mental Health*, *Mental Health Bipolar*, *Reddit Mental Health Data*, *Students Anxiety and Depression*, *Suicidal Mental Health*, *Suicidal Tweet Detection*, and *3k Conversations Dataset for Chatbot*. Sources span forum-style threads, microblog posts, and short dialogues; we mapped them into one table with a single text column per example and a shared seven-way label set. All splits were passed through the same preprocessing and augmentation pipeline—lowercasing; stripping brackets, hyperlinks, HTML or special tags, punctuation, and redundant whitespace—followed by back-translation (English–French–English with TextBlob) and a second pass through the same cleaner so augmented lines match the rest of the corpus. Classical models consume TF–IDF vectors, CNN and RNN baselines use pre-trained Word2Vec embeddings, and transformer runs use the respective subword tokenizers, as summarized in the Preprocessing list below and implemented in Sect. 5. Where class frequencies skewed the metrics, we applied resampling for classical baselines and imbalance-aware losses for deep models, as noted later in the paper.

The final version of the dataset consists of unique person IDs, their statements obtained from conversations, chat logs, and other text-based platforms,

and the corresponding mental health labels. The classification includes the following seven mental health categories:

- **Normal**
- **Depression**
- **Suicidal**
- **Anxiety**
- **Stress**
- **Bipolar**
- **Personal Disorder**

Each sample consists of one or a few sentences describing an individual's mental state. Representative excerpts are included in our public repository.

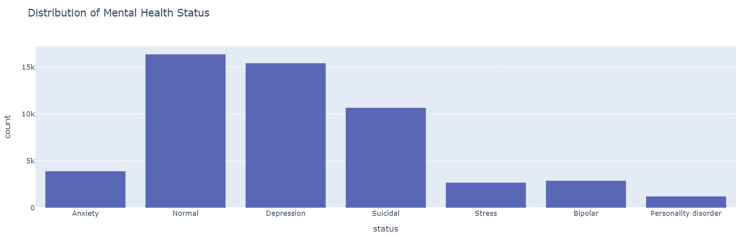


Fig. 1. Distribution of classes in the final dataset.

As shown in Fig. 1, the dataset is imbalanced, with Normal and Depression overrepresented. We mitigate this using imbalance-aware losses: Focal Loss [11], Dice Loss [14], and Tversky Loss [18]; for classical ML baselines we additionally test under-/over-sampling. We preserve proportional class distributions when splitting train/test and use k-fold cross-validation for robust evaluation.

Figure 2 shows that most samples are relatively short, consistent with conversational mental-health text.

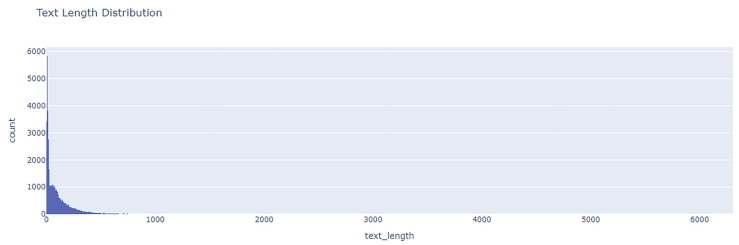


Fig. 2. Text Length distribution of the final dataset.

The combined and refined dataset, along with all the corresponding data, embeddings and TF-IDF vectors, are publicly available and the links for these sources can be found on our GitHub repository.

Preprocessing

- Text normalization: lowercase; remove brackets, hyperlinks, HTML/special tags, punctuation, and redundant whitespace.
- Data augmentation: back-translation EN→FR→EN using TextBlob.
- Representations: TF-IDF for classical ML; pre-trained Word2Vec embeddings [13] for CNN, RNN, LSTM, and GRU baselines; and the corresponding subword tokenizers for the transformer encoders (BERT, ALBERT, DistilBERT, RoBERTa).
- Class imbalance: optional under-/over-sampling for classical baselines and imbalance-aware losses [11,14,18] for deep models.

4 Method

In this section, we describe the methodology used to transfer knowledge from the teacher model to the student model, as well as the architecture and loss functions employed in the DualLatentGNet framework. Our approach leverages a teacher-student knowledge transfer mechanism, where the teacher model, trained with a CNN backbone and two decoders, provides guidance to the student model. This guidance is based on the teacher's feature vectors, which are passed through a Gaussian Mixture Model (GMM) to create a probability distribution. The student model is trained to match the teacher's feature distribution by minimizing the Euclidean distance and the difference in predicted Gaussian components. Additionally, we introduce a loss term $\mathcal{L}_{\text{latentG}}$ that combines the probability density function and Euclidean distance to shape student updates. We also use Focal, Tversky, and Dice losses as alternatives to cross-entropy for class imbalance during training.

Algorithm Explanation

The knowledge transfer is performed in three steps. First, we build teacher feature vectors by concatenating the teacher latent embeddings and prediction logits and fit a GMM over this feature space. Second, for each input, the student computes its feature vector and obtains (i) the most-likely GMM component PDF value and (ii) the Euclidean distance to the teacher feature representation. Finally, these stored quantities guide the student optimization through the latentG loss.

Units of the Architecture

The CNN backbone produces latent vectors that feed two decoders: a classification head and a reconstruction decoder. Training with both objectives shapes a discriminative latent space (see Fig. 3).

Algorithm 1: Teacher-Student Knowledge Transfer Algorithm

Input: Teacher feature vectors: `teacher_latent_v`, `teacher_pred_logits`; dataset (X, y)

Output: Student model predictions and stored PDFs and Euclidean distances.

```
teacher_feature_vectors ← concat[teacher_latent_v, teacher_pred_logits];
gmm ← FitGaussianMixtureModel(teacher_feature_vectors);
Initialize the Student Model;
for each (X, y) in dataset do
    student_feature_vector ← StudentModel(X);
    dist ← EuclideanDistance(student_feature_vector, teacher_feature_vectors);
    pred_pdf ← gmm.predict(student_feature_vector);
    pdf ← MostLikelyGaussian(pred_pdf);
    Store pdf, dist;
end
```

Intuition Behind the Method

The teacher is trained with a dual objective (autoencoding + classification) so that latent representations capture class-discriminative structure (see Fig. 3). We then concatenate teacher latent vectors and logits and fit a GMM to model the feature-space distribution across classes. During student training, the GMM provides the most-likely Gaussian component for each input; we compute both the component PDF value and the Euclidean distance between student and teacher representations, and use them to form the latentG-guided loss alongside the task objectives.

After calculating the pdf and euclid distance, we can denote our latentG (latentGaussian) term as follows:

$$\mathcal{L}_{\text{latentG}} = \alpha\,(1 - p) + \beta \cdot \text{Dist}_{\text{euclid}}$$

where α and β scale the two terms derived from the component PDF complement $(1 - p)$ and the Euclidean distance between teacher/student logits.

So the Dual architecture's total loss will be:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} \times \left(1 + \frac{e}{E} \times \mathcal{L}_{\text{latentG}}\right) + \text{MSE} \times \gamma$$

where e is the current epoch, and E is the total number of epochs. $\mathcal{L}_{\text{cls}}$ is the classification loss, and MSE is used for reconstruction; γ balances the two losses.

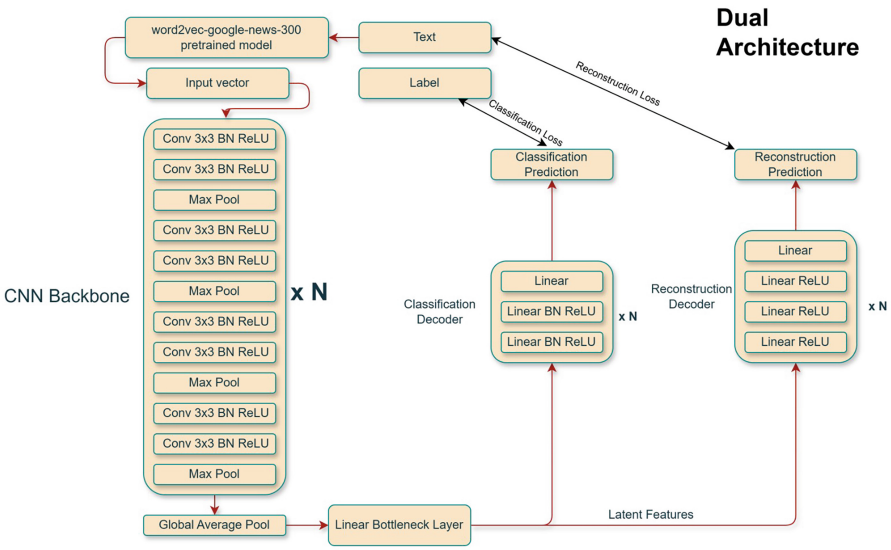


Fig. 3. Dual Architecture Diagram.

In imbalanced settings, we replace $\mathcal{L}_{\text{cls}}$ with Focal, Dice, or Tversky losses [11,14,18].

5 Experiments

This section summarizes how we instantiated and compared models on the merged benchmark. We trained classical TF–IDF pipelines, Word2Vec-based convolutional and recurrent networks, fine-tuned transformer encoders, and our proposed DualLatentGNet under identical splits and the preprocessing choices in Sect. 3. All scripts, notebooks, and hyperparameter configurations are archived in the GitHub repository linked in the introduction.

5.1 Machine Learning Models with TF-IDF Vectors

Using TF-IDF vectors, we train classical baselines (Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest, KNN, XGBoost, and Gradient Boosting). For each model, hyperparameters are selected via grid search with 3-fold cross-validation, and the resulting feature/class correlations motivate the latent-alignment idea used by DualLatentGNet.

5.2 Initial Deep Learning Experiments with Word2Vec Embeddings

We evaluate CNN and sequential baselines with Word2Vec embeddings: the cleaned text is mapped to 300-dim Word2Vec vectors and trained with TextCNN,

along with a lightweight autoencoder for latent extraction. We also train RNN, LSTM, and GRU models using a learnable embedding layer. These baselines capture temporal patterns, but their performance remains lower than the proposed DualLatentGNet and transformer variants.

5.3 Dual Architecture and Latent Representation Alignment

We train a teacher with a dual (autoencoding + classification) objective to shape discriminative latent vectors. We then fit a GMM on the concatenated teacher latent vectors and logits, and use the component PDF complement together with the Euclidean distance between teacher and student representations to form the latentG-guided student objective.

5.4 Transformer-Based Models

We fine-tune transformer baselines (BERT [3], ALBERT [10], DistilBERT [19], and RoBERTa [12]) using their respective pre-trained tokenizers.

5.5 Comparative Framework

DualLatentGNet is compared against classical ML baselines, CNN/RNN deep models, and transformer methods; metrics are reported in the next section.

6 Results

We report accuracy, precision, recall, and F1 on one held-out test split so that numbers remain comparable across model families. For each architecture we ran hyperparameter search and kept the best test result for that family; Table 1 collects these maxima. The training notebooks and configuration files used to produce the table are available on GitHub.

Table 1 summarizes the main numbers across all model families. Within the TF–IDF block, decision trees and random forests reach the low 93% accuracy range, SVM and logistic regression sit in the high 80s to low 90s, while naive Bayes, KNN, and several gradient-boosted configurations fall lower. Among recurrent baselines with trainable embeddings, LSTM reaches 94.14% accuracy and GRU 93.63%, with the vanilla RNN trailing at 88.36%. The fine-tuned transformer block is consistently strong: BERT peaks at 94.83% accuracy, with DistilBERT at 94.43%, RoBERTa at 93.38%, and ALBERT at 92.40%. The CNN-based block shows TextCNN at 91.41% and the dual-objective CNN teacher (DualTextCNN) at 93.01%. **DualLatentGNet** attains the highest scores in the table at **95.09%** accuracy and **95.10%** F1, with precision 95.28% and recall 95.09%, i.e., about 0.26% points above the strongest BERT baseline in absolute terms.

Table 1. Classification performance comparison of different models.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.80	87.00	87.00	87.00
Naive Bayes	71.88	74.00	72.00	72.00
SVM	88.52	88.53	89.34	88.70
Decision Tree	92.25	92.00	92.00	92.00
Random Forest	93.41	94.00	93.00	93.00
KNN	39.35	59.00	39.00	30.00
XGBoost	74.76	75.00	75.00	74.00
Gradient Boosting	77.61	78.00	78.00	77.00
RNN	88.36	86.18	83.72	84.90
LSTM	94.14	93.61	92.06	92.81
GRU	93.63	93.43	91.29	92.30
BERT (Finetuned)	94.83	94.81	94.83	94.80
DistilBERT (Finetuned)	94.43	94.42	94.43	94.42
ALBERT (Finetuned)	92.40	92.39	92.40	92.37
RoBERTa (Finetuned)	93.38	93.37	93.38	93.37
TextCNN	91.41	91.43	91.41	91.38
DualTextCNN (Teacher)	93.01	92.99	93.01	92.97
DualLatentGNet (Proposed)	**95.09**	**95.28**	**95.09**	**95.10**

Table 2. Classification performance comparison of different loss functions.

Model	Loss	Accuracy	Precision	Recall	F1-Score
DualLatentGNet	Dice	94.30	94.43	94.41	94.42
DualLatentGNet	Tversky	94.56	94.75	94.22	94.49
DualLatentGNet	CE	**95.09**	**95.28**	**95.09**	**95.10**

1 CE: Cross Entropy

Table 2 reports an ablation in which the cross-entropy classification term inside DualLatentGNet is replaced by Dice or Tversky loss. Cross-entropy retains the highest accuracy at 95.09%, while Tversky reaches 94.56% and Dice 94.30%; the same ordering holds for precision, recall, and F1.

7 Discussion

Where the Gain Comes from. The 0.26-point margin of DualLatentGNet over the best BERT baseline is small in absolute terms but consistent across precision, recall, and F1 (Table 1). Because the only difference between the proposed model and its CNN teacher (DualTextCNN, 93.01%) is the addition of the latentG-guided student objective, the roughly two-point lift over the teacher and the

0.26-point lift over BERT are most readily attributed to that mechanism rather than to the underlying CNN backbone. We read this as evidence that explicit alignment to the teacher's latent geometry contributes signal that contextual token embeddings alone do not fully capture on this aggregate benchmark, but not as a claim that contextual modeling should be abandoned. The two views are complementary, and combining a transformer encoder with the latentG term is a natural next step.

Loss-Function Ablation. The ranking in Table 2 is informative even though the differences are modest. On highly imbalanced segmentation problems Dice and Tversky usually outperform cross-entropy precisely because they downweight easy majority pixels; here the opposite holds. A plausible reading is that latentG-guided alignment already reshapes the student's feature space so that minority-class samples are pulled towards the right Gaussian component, and the standard classification term then suffices. Adding Dice or Tversky on top appears to interact unfavorably with that geometric pressure on this split, rather than complementing it.

Recurrent vs. Contextual Baselines. The vanilla RNN lags the gated variants noticeably (Table 1), in line with the well-known difficulty of retaining information over variable-length posts with simple recurrence. LSTM slightly edges GRU here, plausibly because explicit memory cells help on mixed-length snippets. Even without pre-trained contextual embeddings, trainable embeddings combined with gated recurrence reach the low 94% range, which justifies keeping these models in the comparison alongside transformers and indicates how much of the headroom on this benchmark is reachable without large-scale pre-training.

Cost of the Proposed Method. A reviewer noted that the gain over BERT comes with a complexity increase, and we want to be explicit about it. Training Dual-LatentGNet requires three steps that fine-tuned BERT does not: (i) training the dual-objective CNN teacher with both classification and reconstruction heads, (ii) fitting a Gaussian mixture model on the concatenated teacher latent vectors and logits, and (iii) training the student under the additional latentG term, which involves a per-batch lookup of the most-likely Gaussian component and a Euclidean-distance computation against teacher representations. At inference time, however, only the student CNN runs, so deployment cost is on the same order as a single CNN forward pass and is markedly lower than running a fine-tuned BERT-base. The trade-off is therefore extra training cost and an additional fitted GMM in exchange for cheaper inference and the observed accuracy gain.

Threats to Validity. High offline accuracy on social-media text does not establish clinical utility. Mental-health categories such as *Suicidal* and *Personal Disorder* carry consequences that no aggregate accuracy number captures: false negatives matter more than false positives, calibration on the deployment population matters more than test-set ranking, and fairness across demographics matters more

than overall metrics. The seven categories in our label schema are also defined at a granularity that mixes diagnostic-style labels (e.g., *Bipolar*) with state-style labels (e.g., *Stress*); an inferred class is therefore best read as a routing cue toward human review rather than as a diagnosis. We also report point metrics on a single held-out split; reporting variance across seeds and folds, and uncertainty estimates per prediction, would strengthen the evidence base.

8 Conclusion

This paper assembled a seven-way benchmark by harmonizing nine public mental-health text sources under a single preprocessing and augmentation pipeline, compared classical TF–IDF models, CNN/RNN/LSTM/GRU baselines, and fine-tuned transformer encoders, and introduced DualLatentGNet, a teacher–student framework in which a Gaussian mixture is fit to the teacher's concatenated latent vectors and logits and a latentG term ties the student to that geometry alongside the standard classification loss. Under shared splits, DualLatentGNet reached 95.09% accuracy, narrowly above the strongest fine-tuned BERT baseline at 94.83% and about two points above its own CNN teacher at 93.01%, with the gain consistent across precision, recall, and F1.

We do not read this margin as a claim that contextual models should be replaced. The point of interest is that an explicit alignment to a teacher's latent geometry, in addition to its predictions, can carry signal that contextual token embeddings alone do not fully capture on this benchmark, while keeping inference cost on the order of a single CNN forward pass. The most informative next experiments are, in our view, an integration of pre-trained encoders into the student or teacher with the latentG term retained; reporting variance across seeds and folds together with calibration and per-prediction uncertainty estimates; and, for resource-constrained deployment, compression techniques such as pruning or quantization applied to DualLatentGNet (or to its teacher), provided they preserve the behavior observed here. Text classification for mental-health monitoring should remain a triage aid rather than a diagnostic substitute, and the work above is offered in that spirit.

References

1. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ. Couns. (2005). https://doi.org/10.1016/j.pec.2004.09.008
2. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N.: SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING), pp. 1485–1497 (2018). https://aclanthology.org/C18-1126/
3. 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 (2019). https://doi.org/10.18653/v1/N19-1423
4. Drougkas, G., Bakker, E.M., Spruit, M.: Multimodal machine learning for language and speech markers identification in mental health. BMC Med. Inform. Decis. Mak. (2024). https://doi.org/10.1186/s12911-024-02772-0
5. European Commission: European pact for mental health and well-being (2008). https://health.ec.europa.eu/publications/european-pact-mental-health-and-well-being_en
6. Fatemi, M., El Asri, L., Schulz, H., He, J., Suleman, K.: Policy networks with two-stage training for dialogue systems. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2016). https://doi.org/10.18653/v1/W16-3613, https://aclanthology.org/W16-3613/
7. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Mental Health (2017). https://doi.org/10.2196/mental.7785
8. Insel, T.R., Scolnick, E.M.: Cure therapeutics and strategic prevention: raising the bar for mental health research. Mol. Psychiatry (2006). https://doi.org/10.1038/sj.mp.4001777
9. Kour, H., Gupta, M.K.: An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. Multimedia Tools Appl. (2022). https://doi.org/10.1007/s11042-022-12648-y
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019). https://arxiv.org/abs/1909.11942
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). https://arxiv.org/abs/1708.02002
12. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019). https://arxiv.org/abs/1907.11692
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013). https://arxiv.org/abs/1301.3781
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV) (2016). https://arxiv.org/abs/1606.04797
15. Olfson, M., Marcus, S.C., Druss, B., Pincus, H.A.: National trends in the use of outpatient psychotherapy. Am. J. Psychiatry (2002). https://doi.org/10.1176/appi.ajp.159.11.1914
16. Pietquin, O., Geist, M., Chandramohan, S.: Sample efficient on-line learning of optimal dialogue policies with Kalman temporal differences. In: IJCAI (2011). https://ijcai.org/Abstract/11/314
17. Saleela, D., et al.: BERT-based text classification pipeline for categorising mental health statements. In: Information Systems for Intelligent Systems (2026). https://doi.org/10.1007/978-3-032-12993-2_3, https://link.springer.com/chapter/10.1007/978-3-032-12993-2_3
18. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging (2017).https://doi.org/10.48550/arXiv.1706.05721, https://arxiv.org/abs/1706.05721

19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019). https://arxiv.org/abs/1910.01108
20. Schatzmann, J., Thomson, B., Young, S.: Statistical user simulation with a hidden agenda. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue (2007). https://doi.org/10.18653/v1/2007.sigdial-1.48, https://aclanthology.org/2007.sigdial-1.48/
21. Schoene, A.M., Lacey, G., Turner, A.P., Dethlefs, N.: Dilated LSTM with attention for classification of suicide notes. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pp. 136–145 (2019). https://doi.org/10.18653/v1/D19-6217, https://aclanthology.org/D19-6217/
22. World Health Organization: Mental health action plan 2013-2020 (2013). https://www.who.int/publications/i/item/mental-health-action-plan-2013-2020
23. Young, S., Gasic, M., Thomson, B., Williams, J.D.: POMDP-based statistical spoken dialogue systems: a review. Proc. IEEE (2013). https://doi.org/10.1109/JPROC.2012.2225812

Evaluating Multi-objective Evolutionary Algorithms for Fairness-Aware Decision Tree Tuning

Ana Januário[1(✉)], Maria João Alves[1,2], and Carlos Henggeler Antunes[2,3]

[1] CeBER, Faculty of Economics, University of Coimbra, Coimbra, Portugal
uc2023140299@student.uc.pt, mjalves@fe.uc.pt
[2] Institute of Systems and Computer Engineering (INESC), Coimbra, Portugal
[3] Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal
ch@deec.uc.pt

Abstract. Fairness is increasingly recognised as a critical requirement in machine learning applications, particularly in sensitive domains such as healthcare, where misclassifications may disproportionately affect specific demographic groups. This study investigates how multi-objective hyperparameter optimisation can balance predictive performance and algorithmic fairness in decision tree models. The optimisation model simultaneously minimises classification error and gender-based unfairness, defined as disparities in true positive rates between male and female groups. Three optimisation strategies are compared: an exhaustive grid search and two evolutionary multi-objective algorithms, NSGA-II and MOEA/D. Their resulting solution sets are analysed by comparing Pareto fronts, receiver operating characteristic (ROC) curves and gender-stratified confusion matrices to assess both overall predictive performance and the distribution of classification errors across demographic groups.

Grid search identifies the configuration with the highest area under the ROC curve due to its exhaustive exploration of the hyperparameter space. NSGA-II significantly reduces computational time while retaining competitive solutions, though with slightly lower discriminative capability. MOEA/D discovers configurations that closely match the predictive performance of grid search and display more balanced error distributions across genders. These results show the potential of evolutionary multi-objective optimisation as an effective and practical approach for fairness-aware hyperparameter tuning. They also highlight the importance of assessing group-specific error distributions beyond aggregate fairness metrics to obtain a more reliable and nuanced understanding of model behaviour in sensitive decision-making contexts.

Keywords: Multi-Objective Optimisation · Hyperparameter Optimisation · Machine Learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 274–291, 2027.
https://doi.org/10.1007/978-3-032-30494-0_18

1 Introduction

Decision tree classifiers remain widely used in machine learning due to their intuitive structure, ease of interpretation, and ability to handle heterogeneous data types [4,13]. In their classical formulation, trees recursively partition the feature space into increasingly homogeneous regions with respect to the target variable. At each internal node, a splitting rule is selected to maximise an impurity reduction criterion, while terminal nodes assign class labels based on the majority class of the training samples reaching that node. This hierarchical and rule-based representation makes decision trees particularly attractive in decision support applications, where model transparency and interpretability are essential [2,13].

Despite their conceptual simplicity, the predictive behaviour of decision trees is strongly influenced by a set of hyperparameters that control the tree-growing process. Parameters such as maximum depth, minimum number of samples required to split a node and pruning thresholds determine the complexity of the resulting model and its ability to generalise beyond the training data. Trees grown with weak constraints may capture complex relationships at the risk of overfitting, whereas stricter constraints typically produce simpler models with improved generalisation but reduced flexibility. Consequently, the careful selection of hyperparameter configurations is crucial to achieving an appropriate balance between predictive accuracy, robustness and interpretability [1,13].

Hyperparameter optimisation seeks to identify configurations of the control parameters that maximise model performance according to a chosen evaluation criterion. Traditional approaches, such as manual tuning or exhaustive grid search, evaluate predefined sets of candidate configurations and select the one that optimises a chosen validation metric [1]. Although straightforward to implement, these strategies become computationally expensive as the number of hyperparameters increases and may struggle to capture complex trade-offs when multiple objectives must be considered simultaneously. For this reason, more advanced search strategies based on evolutionary algorithms and other meta-heuristic techniques have been increasingly explored to navigate high-dimensional hyperparameter spaces more efficiently [9,11,15,18].

While most hyperparameter optimisation procedures focus primarily on improving predictive performance, growing attention has been paid to the societal implications of machine learning systems. A substantial body of research has highlighted systematic disparities in model outcomes across demographic groups, bringing algorithmic fairness to the forefront of machine learning research [5,14]. In classification tasks, unfairness may arise when models display different error rates, true positive rates or decision thresholds for protected groups defined by attributes such as gender, race or age [12]. Fairness-aware learning, therefore, seeks to design models that satisfy explicit fairness criteria while maintaining acceptable predictive performance [14,17]. Bias mitigation can be addressed at different stages of the modelling pipeline: through pre-processing techniques that modify the training data, by incorporating fairness constraints directly into the learning algorithm or by post-processing model outputs to reduce disparities [14].

Among the various fairness concepts proposed in the literature, group-based criteria such as equal opportunity, which requires similar true positive rates across protected groups, are particularly relevant in sensitive domains such as healthcare and credit scoring [12].

Recent work highlights the challenges of optimising fairness in practice. The choice of fairness metric is itself non-trivial: Uddin et al. [17] show that the same dataset can yield fair outcomes under one algorithm and unfair outcomes under another, depending on the fairness definition adopted. Corbett-Davies et al. [5] demonstrate that widely used formal fairness criteria often lead to Pareto-dominated decision policies, suggesting that no axiomatic definition of fairness is universally adequate. d'Aloisio et al. [7] show that, even when fairness is explicitly considered as an objective to be optimised together with predictive performance, the resulting trade-offs depend heavily on the choice of the metric and model type. These findings suggest that, rather than enforcing fairness as a post-hoc constraint, it may be more effective to incorporate it directly into the model selection process through hyperparameter optimisation.

These developments motivate treating hyperparameter selection as a multi-objective optimisation problem, in which predictive performance and fairness are pursued simultaneously rather than independently. Instead of collapsing multiple evaluation criteria into a single scalar objective using ad hoc weights, multi-objective optimisation explicitly searches for Pareto-optimal solutions that represent efficient trade-offs between competing goals [8]. Evolutionary multi-objective algorithms, such as NSGA-II and MOEA/D, are especially well suited to this setting. By maintaining a diverse population of candidate solutions, these algorithms can approximate the Pareto front in a single optimisation run, thereby providing information on different trade-offs between objectives [9,18]. Moreover, both algorithms have demonstrated strong empirical performance across a wide range of multi-objective optimisation problems and have become standard benchmarks in the evolutionary optimisation literature. When applied to decision tree hyperparameter optimisation, these algorithms enable practitioners to explore how different configurations balance classification error against group-level fairness measures, such as disparities in true positive rates between gender groups, and to select models that best align with the requirements of a given decision-making context [6,15].

In this work, we investigate the multi-objective hyperparameter optimisation of decision tree classifiers, where predictive performance and fairness are treated as objectives to optimise. Specifically, we jointly minimise classification error and gender-based unfairness, which is measured as disparities in true positive rates between male and female groups according to the equal opportunity criterion [12]. To explore the trade-offs between these objectives, we compare an exhaustive grid search strategy with two evolutionary multi-objective algorithms, NSGA-II and MOEA/D, and analyse the resulting Pareto fronts obtained under each optimisation approach [9,15,18]. The proposed framework aims to provide a systematic approach for examining how different hyperparameter configura-

tions affect the trade-off between predictive accuracy and group-level fairness in decision tree models.

The main contributions of this work are:

- a multi-objective framework for fairness-aware hyperparameter tuning;
- a comparative analysis of grid search and well-established evolutionary optimisation methods;
- a detailed group-level error analysis using confusion matrices.

2 Methodology

In the framework of a decision tree classifier for predicting heart disease, this study develops a model selection structure formulated as a multi-objective hyperparameter optimisation problem. Different decision tree hyperparameter configurations are generated and explored using grid search and evolutionary search strategies to optimise predictive performance and fairness with respect to gender.

To ensure rigorous evaluation, the dataset was randomly partitioned into three mutually exclusive subsets: a training set used for model fitting, a validation set for identification of Pareto-optimal solutions for hyperparameter tuning and an independent test set for final performance assessment.

Hyperparameter optimisation was carried out using three distinct approaches: exhaustive Grid Search, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [8,9] and the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [18], enabling a comparative analysis of their efficiency and solution quality.

The hyperparameter search space has been designed to explore the trade-off between model complexity and predictive performance. It includes the following parameters:

- Maximum tree depth (max_depth): integer values from 1 to 20.
- Minimum number of samples required to split a node (min_samples_split): integer values from 2 to 50.
- Maximum number of features considered for each node split (max_features): integer values from 1 to 7.

2.1 Performance Metrics

Model predictions were used to compute several performance metrics. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

Precision and recall are defined in (1) and (2):

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

$$TPR(Recall) = \frac{TP}{TP + FN} \tag{2}$$

Recall is also referred to as the true positive rate (TPR), as it measures the proportion of actual positive instances that are correctly identified by the model. These two measures are combined in the $F1_{score}$ ($F1$), which corresponds to the harmonic mean of precision and recall, as in (3):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

The classification error (E) considered in this study is defined in (4):

$$E = 1 - F1 \tag{4}$$

To evaluate fairness considering groups, the recall was computed separately for male and female individuals by partitioning the validation and test set according to the `Sex` attribute, thereby enabling the assessment of group-specific diagnostic performance conditional on the presence of heart disease. The unfairness metric (U) was defined as the absolute difference between the true positive rates of the two groups, as in (5):

$$U = |TPR_{men} - TPR_{women}| \tag{5}$$

where TPR_{men} and TPR_{women} denote the recall scores for male and female individuals, respectively.

2.2 Multi-objective Framework

Let $\mathcal{M}$ be the set of all trained models. Model selection was based on Pareto dominance, considering classification error (E) and unfairness (U) as the two objectives to be minimised. A model $M_i \in \mathcal{M}$ is said to dominate another model $M_j \in \mathcal{M}$ if it is not worse than M_j with respect to both objectives and strictly better in at least one of them, that is, with at least one strict inequality (6).

$$E_i \leq E_j \text{ and } U_i \leq U_j \tag{6}$$

The goal is, therefore, to identify the Pareto front, containing the models that offer the best possible trade-offs, where any further reduction in unfairness would inevitably result in a higher classification error and vice versa. In case of ties, where multiple configurations yielded identical error and unfairness values, the Occam's Razor principle was applied, giving preference to the least complex model [2]. Each solution represents a specific decision tree hyperparameter configuration, trained on the training set and evaluated on the validation set to obtain its classification error and unfairness values.

To assess the generalisation performance of the Pareto-optimal solutions, all models identified as Pareto optimal during the validation phase were evaluated

on an independent test set. We further rank Pareto-optimal models using the Area Under the ROC Curve (AUC), which is not an optimisation objective but provides an additional discriminative evaluation. ROC curves were generated for the five models exhibiting the highest AUC values. ROC curves illustrate the relationship between the true positive rate and the false positive rate across all possible decision thresholds, providing a threshold-independent evaluation of model performance. The AUC summarises this relationship into a single scalar value that reflects the model's overall ability to discriminate between positive and negative instances. Therefore, these indicators provide a comprehensive assessment of discriminative performance and support the final model selection [10].

The solutions produced by NSGA-II and MOEA/D were compared to the non-dominated solutions obtained with the exhaustive grid search. For selected models from all optimisation strategies, confusion matrices disaggregated by gender were analysed to examine error distributions across demographic groups and verify fairness properties in greater detail.

2.3 Dataset and Implementation Options

The analysis employed a dataset about personal health indicators associated with heart disease: heart_2020_cleaned dataset, publicly available on the Kaggle platform[1]. It includes a binary target variable, `Heart Disease`, denoting the presence or absence of heart disease, and a set of explanatory variables encompassing demographic, behavioural, and health-related attributes (e.g., `BMI` (body mass index), `Smoking`, `Alcohol Drinking`, `Stroke`, `Physical Health`, `Mental Health`, `Difficulty Walking`, `Sex`, `Age Category`, `Race`, `Diabetic`, `Physical Activity`, `General Health`, `Sleep Time`, `Asthma`, `Kidney Disease`, and `Skin Cancer`).

The target variable is imbalanced: approximately 91.5% of instances correspond to individuals without heart disease, whereas 8.5% represent diagnosed cases. Consequently, accuracy was deemed unsuitable as it may mask poor predictive performance for the minority class. Instead, the $F1_{score}$ was adopted as the primary performance metric, given its robustness to class imbalance and its formulation as the harmonic mean of precision and recall [2,10].

For grid search, a decision tree model was trained for each combination of hyperparameters (max_depth, min_samples_split, max_features), resulting in 6,860 distinct models being fitted on the training set. Each trained model was subsequently evaluated on the validation set.

The evolutionary multi-objective optimisation approaches were implemented using the same decision space. Both NSGA-II [9] and MOEA/D [18] were configured with identical parameter settings to enable a fair comparison. The initial population was generated using an integer random sampling scheme. Variation operators consisted of simulated binary crossover (SBX) with probability

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

$p_c = 0.9$ and distribution index $\eta_c = 15$, and polynomial mutation with probability $p_m = 1/3 \approx 0.33$ and distribution index $\eta_m = 20$. All offspring solutions were repaired using a rounding operator to ensure integer feasibility.

For NSGA-II, the population size was set to $N = 50$, and the algorithm was executed for 60 generations, using non-dominated sorting and crowding distance to promote convergence towards and diversity along the Pareto front [9]. Further experiments with a larger number of generations produced similar results, indicating that the chosen configuration was sufficient to reach stable solutions.

The same termination criterion (60 generations) and population size ($N = 50$) were adopted for MOEA/D. In this case, the algorithm decomposes the bi-objective problem into a set of scalar subproblems defined by uniformly distributed reference directions and employees neighbourhood-based mating. The specific parameters were set to $T = 15$ neighbours and probability of selecting parents from the neighbourhood $p_{\text{neigh}} = 0.7$.

All analyses were conducted in a Python environment. The optimisation algorithms were implemented using the pymoo framework [3] and the decision tree classifier was implemented using the scikit-learn library[2].

3 Results

The exhaustive grid-search procedure evaluated a total of 6,860 candidate models, requiring approximately 20 min of training time. In contrast, NSGA-II evaluated 3,000 candidate solutions in approximately 6 min and 17 s, while MOEA/D assessed 3,060 candidate solutions in 19 min and 6 s.

The Pareto front obtained from the grid search on the validation set includes 15 nondominated solutions, revealing a clear trade-off between predictive performance and fairness (Fig. 1). In this dataset, models with lower classification error tend to exhibit higher unfairness, whereas solutions achieving near-zero unfairness require sacrificing predictive accuracy. Table 1 reports the Pareto-optimal configurations and their objective values. For instance, model G1 attains the lowest error (0.7675) but the highest unfairness (0.0285), while model G15 achieves perfect unfairness at the cost of substantially higher classification error (0.8882).

NSGA-II also produced 15 non-dominated solutions, recovering 9 of the 15 Pareto-optimal configurations identified by the exhaustive grid search (60%). MOEA/D identified 10 Pareto-optimal solutions, recovering 5 of the Pareto-optimal configurations obtained by grid search (50%). Although MOEA/D generated fewer Pareto-optimal solutions, several of its configurations lie close to the Pareto front of the grid search, indicating competitive performance in approximating the trade-off surface. Figure 2 shows the fronts generated by NSGA-II and MOEA/D on the validation set.

Table 2, compares validation and test performance for the grid-search Pareto-optimal configurations. The classification error on the test set remains relatively stable across datasets, ranging from 0.679 to 0.728. In several cases, the test

[2] https://scikit-learn.org.

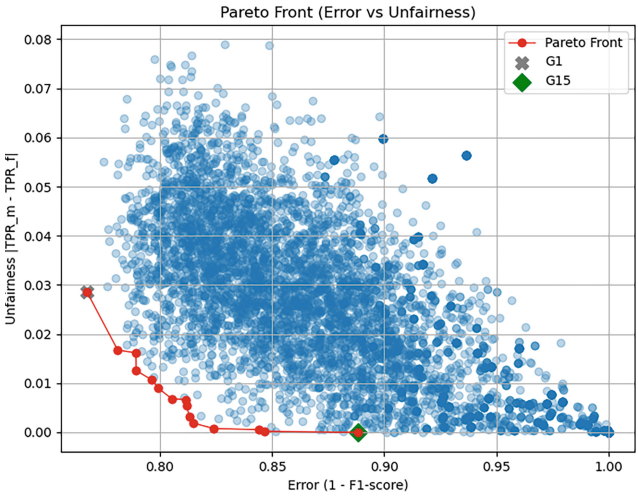


Fig. 1. Pareto front obtained on the validation set, illustrating the trade-off between classification error and gender-based unfairness across decision tree hyperparameter configurations.

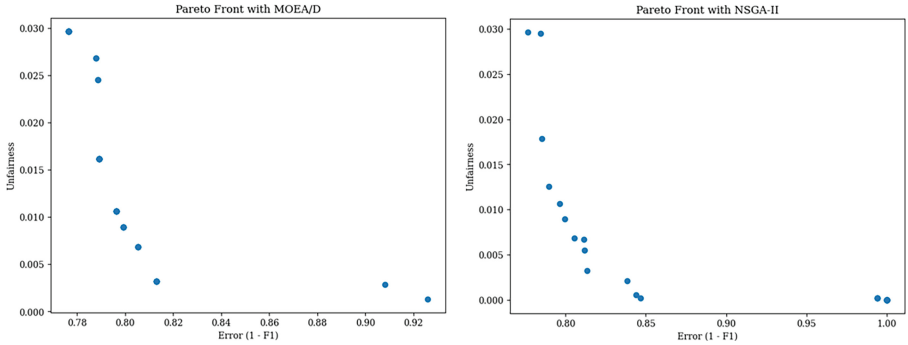


Fig. 2. Pareto front obtained on the validation set for MOEA/D and NSGA-II, illustrating the trade-off between classification error and unfairness across decision tree hyperparameter configurations.

error is even lower than the corresponding validation error, suggesting that the models do not suffer from severe overfitting and display strong generalisation capability.

In contrast, fairness metrics exhibit a markedly different behaviour. Although the validation set unfairness ranges from 0.000 to 0.029, test set unfairness increases substantially, typically falling between 0.09 and 0.20. Even configurations that achieve near-zero unfairness during validation display noticeable disparities on unseen data. This pattern indicates that fairness metrics may be considerably more sensitive to distributional differences between datasets than

Table 1. Pareto-optimal solutions obtained via Grid Search (validation set).

Model	Error	Unfairness	Depth	min_samples_split	max_features
G1	0.7675	0.0285	20	6	7
G2	0.7812	0.0167	19	7	7
G3	0.7893	0.0162	19	2	7
G4	0.7894	0.0125	18	4	6
G5	0.7965	0.0106	19	2	3
G6	0.7994	0.0089	19	2	4
G7	0.8052	0.0068	20	2	2
G8	0.8113	0.0066	17	7	6
G9	0.8119	0.0055	20	4	2
G10	0.8131	0.0032	20	2	1
G11	0.8147	0.0019	15	7	6
G12	0.8241	0.0007	13	15	6
G13	0.8440	0.0005	19	10	1
G14	0.8467	0.0002	19	5	1
G15	0.8882	0.0000	10	9	5

conventional predictive metrics, implying that fairness-optimised configurations may not generalise reliably.

A similar pattern is observed for NSGA-II. As shown in Table 3, test set error remains within a narrow range (0.689 to 0.797), while fairness deteriorates substantially when evaluated out-of-sample. Test set unfairness commonly lies between 0.10 and 0.20, confirming that fairness estimates obtained during optimisation may underestimate true disparities.

The results obtained with MOEA/D follow the same general trend (Table 4). Test set classification error remains consistent with results from grid search and NSGA-II (0.696 to 0.728), whereas fairness again degrades when evaluated out-of-sample. Although validation unfairness is generally below 0.03, the corresponding test values increase substantially for most configurations. Most MOEA/D configurations show test set unfairness above 0.15, suggesting that fairness generalisation issues are persistent across optimisation strategies. Overall, the MOEA/D solutions exhibit comparable predictive performance and fairness behaviour to those obtained with NSGA-II and the grid search.

3.1 Analysis Based on Area Under the ROC Curve

To further identify the most suitable model among the Pareto-optimal solutions, we evaluated the Area Under the ROC Curve on the test set. AUC complements the bi-objective error and fairness evaluation by providing a threshold-independent measure of discriminative ability. Figure 3 presents the ROC curves

Table 2. Results comparison: validation vs. test sets for Pareto-optimal models obtained by the grid-search.

Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Features
0.8241	0.6792	0.0008	0.1436	13	15	6
0.8882	0.6799	0.0000	0.0948	10	9	5
0.8148	0.6855	0.0019	0.1552	15	7	6
0.8114	0.6894	0.0067	0.1594	17	7	6
0.7894	0.6986	0.0126	0.1832	18	4	6
0.7994	0.6999	0.0089	0.1722	19	2	4
0.8440	0.7000	0.0006	0.1481	19	10	1
0.7675	0.7038	0.0285	0.1778	20	6	7
0.7893	0.7053	0.0162	0.1739	19	2	7
0.7812	0.7058	0.0167	0.1781	19	7	7
0.7965	0.7066	0.0107	0.1690	19	2	3
0.8052	0.7176	0.0068	0.1971	20	2	2
0.8467	0.7189	0.0002	0.1484	19	5	1
0.8119	0.7234	0.0055	0.1672	20	4	2
0.8132	0.7281	0.0032	0.1951	20	2	1

Table 3. Results comparison: validation vs. test sets for NSGA-II

Model	Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Features
N1	0.8114	0.6894	0.0066	0.1594	17	7	6
N2	0.7843	0.6962	0.0295	0.1697	20	13	6
N3	0.7894	0.6986	0.0125	0.1832	18	4	6
N4	0.7994	0.6999	0.0089	0.1722	19	2	4
N5	0.8440	0.7000	0.0006	0.1481	19	10	1
N6	0.7854	0.7004	0.0179	0.1749	20	9	6
N7	0.7765	0.7065	0.0297	0.1732	20	2	5
N8	0.7964	0.7066	0.0106	0.1689	19	2	3
N9	0.8382	0.7098	0.0021	0.1284	19	7	2
N10	0.8052	0.7176	0.0068	0.1971	20	2	2
N11	0.8467	0.7189	0.0002	0.1484	19	5	1
N12	0.8119	0.7234	0.0055	0.1672	20	4	2
N13	0.8131	0.7281	0.0032	0.1951	20	2	1
N14	0.9941	0.7603	0.0002	0.0293	3	9	3
N15	1.0000	0.7975	0.0000	0.1006	1	17	1

Table 4. Results comparison: validation vs. test sets for MOEA/D

Model	Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Max features
M1	0.926	0.696	0.001	0.038	8	9	3
M2	0.799	0.700	0.009	0.172	19	2	4
M3	0.788	0.700	0.027	0.183	19	2	5
M4	0.789	0.702	0.025	0.159	20	5	5
M5	0.908	0.705	0.003	0.088	8	9	4
M6	0.789	0.705	0.016	0.174	19	2	7
M7	0.777	0.707	0.030	0.173	20	2	5
M8	0.796	0.707	0.011	0.169	19	2	3
M9	0.805	0.718	0.007	0.197	20	2	2
M10	0.813	0.728	0.003	0.195	20	2	1

for the five Pareto-optimal grid-search configurations with the highest AUC values, and Table 5 summarises their performance.

Table 5. Top 5 Pareto-optimal models obtained by the grid-search ranked by global AUC (test set)

Rank	Depth	Samples	Features	AUC	Error	Unfairness
1	10	9	5	0.8116	0.6799	0.0948
2	13	15	6	0.7877	0.6792	0.1436
3	15	7	6	0.7505	0.6855	0.1552
4	19	10	1	0.7239	0.7000	0.1481
5	17	7	6	0.7129	0.6894	0.1594

The best-performing model in the test set under grid search (depth = 10, min_samples_split = 9, and max_features = 5) achieved the highest AUC (0.8116), a low classification error (0.6799) and the lowest unfairness (0.0948) among the top five models.

A similar AUC-based evaluation was performed for NSGA-II in the test set (Fig. 4). The best configuration achieved an AUC of 0.757, corresponding to a relatively shallow decision tree (depth = 3, min_samples_split = 9, max_features = 3). Although this model exhibits a moderate level of discriminative ability, its AUC remains lower than that obtained by the model identified through grid search. This result suggests that the configuration found by this method prioritised solutions with lower unfairness rather than maximising the AUC.

For MOEA/D, the best configuration achieved an AUC of 0.8086 (depth = 8, min_samples_split = 9, and max_features = 3), which is very close to the best AUC obtained with grid search (0.8116), Fig. 5. Although MOEA/D required more computation time than NSGA II, and even approached that of

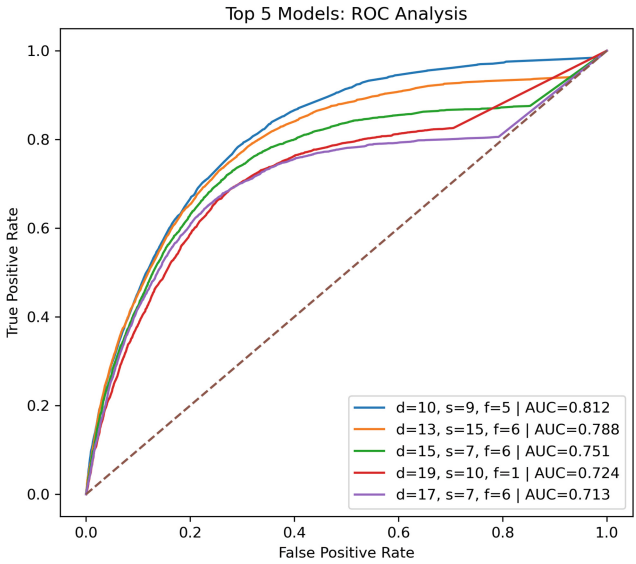


Fig. 3. ROC curves of top-performing models sorted by AUC in the test set for grid search.

grid-search, its top-performing solution demonstrates that decomposition-based evolutionary optimisation can effectively locate high-quality regions of the Pareto front. Unlike NSGA-II, which relies on global non-dominated sorting and crowding distance mechanisms, MOEA/D decomposes the multi-objective problem into multiple scalar subproblems and performs neighbourhood- based updates. The management of these neighbourhood structures may introduce additional computational overhead during the optimisation process.

An interesting pattern was observed when studying the non-dominance considering the classification error and unfairness in the test set. For the solutions obtained with grid search and MOEA/D, we noticed that non-dominated solutions were also the configurations corresponding to the highest AUC values, thus suggesting an alignment between Pareto-optimality in error/unfairness space and overall discriminative performance assessed by AUC. For NSGA-II, three of the five solutions selected for AUC evaluation were non-dominated, including the configuration that achieved the highest AUC.

When examining all non-dominated solutions jointly across methods in the test set, the configuration with the lowest classification error was obtained via grid search, whereas the solution with the lowest unfairness was obtained with NSGA-II (Fig. 6). The solutions obtained by MOEA/D, however, tend to occupy intermediate regions of the Pareto front, suggesting a more balanced trade-off between predictive performance and fairness. Such patterns may also reflect the complex relationship between predictive performance and fairness, where the extent of the trade-off can vary depending on dataset characteristics and on the specific fairness metric considered.

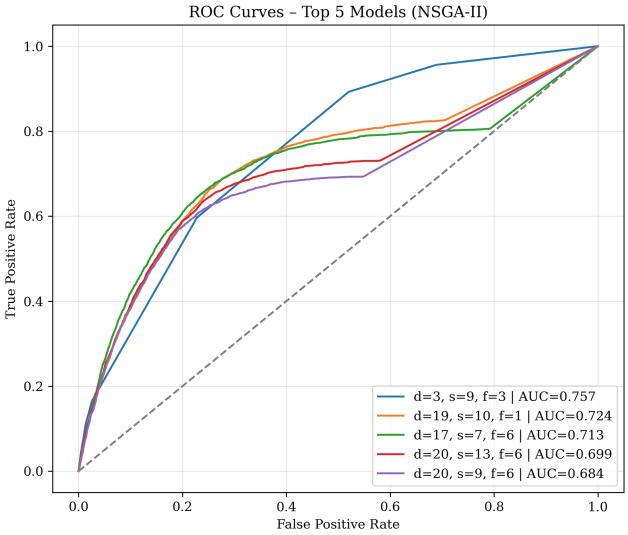


Fig. 4. ROC curves of top-performing models sorted by AUC in the test set during NSGA-II.

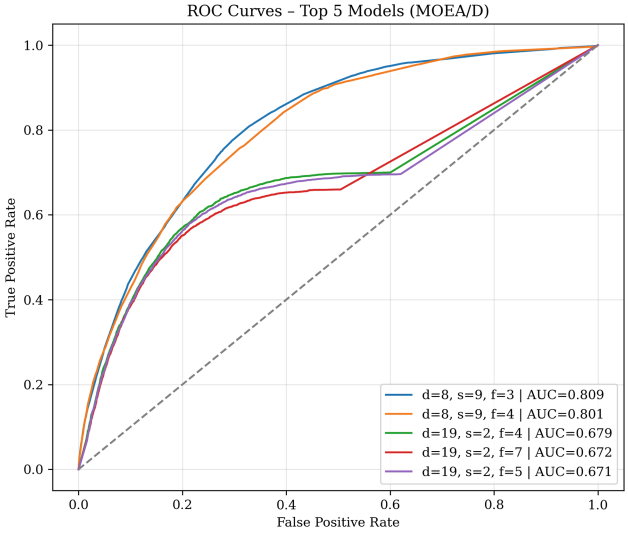


Fig. 5. ROC curves of top-performing models sorted by AUC in the test set during MOEA/D.

3.2 Confusion-Matrix Analysis

While the aggregate unfairness metric provides an overall indication of disparity between demographic groups, it does not reveal how prediction errors are distributed across these groups. The group level confusion-matrix analysis provides

further insights into fairness behaviour. In particular, the False Positive Rate (FPR) is informative because it measures how frequently individuals without the condition are incorrectly classified as positive within each group. Differences in FPR may indicate unequal exposure to unnecessary diagnostic procedures, making this metric especially relevant in clinical decision-support settings.

The FPR is defined in (7):

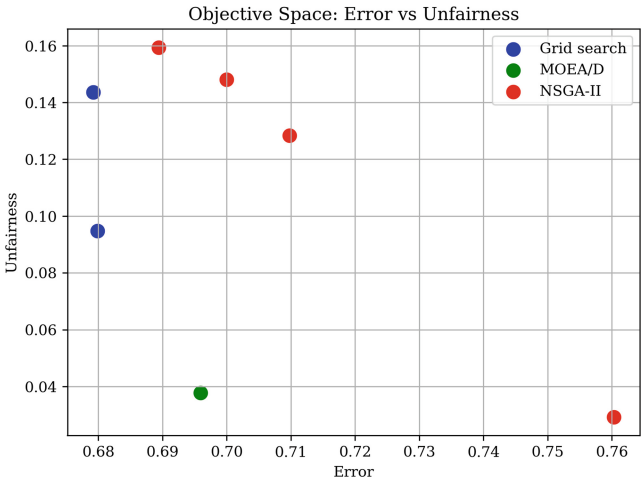


Fig. 6. Non-dominated solutions of each method in the test set.

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

For the best grid search configuration (Fig. 7), recall for men (0.8212) exceeds that for women (0.7264), indicating women are more likely to receive false negatives. Conversely, men exhibit a higher false positive rate (0.3368 vs. 0.2500), implying greater exposure to unnecessary follow-up procedures.

For the best NSGA-II configuration (Fig. 8), recall is substantially higher for both groups (0.9048 for men and 0.8755 for women), but this improvement is accompanied by a significant increase in false positive rates (0.4848 for men and 0.5509 for women). This indicates that the model tends to classify a large fraction of observations as positive, increasing sensitivity at the expense of predictive reliability.

The best MOEA/D configuration (Fig. 9) produces a more balanced distribution of errors: recall values remain competitive (0.8244 for men, 0.7865 for women), and false positive rates are nearly identical across groups (0.3286 for men and 0.3285 for women). This suggests that MOEA/D yields a more symmetric error distribution while maintaining strong discriminative performance.

Overall, the group-level analysis highlights important differences in how each optimisation strategy balances accuracy and fairness. Grid-search achieves the

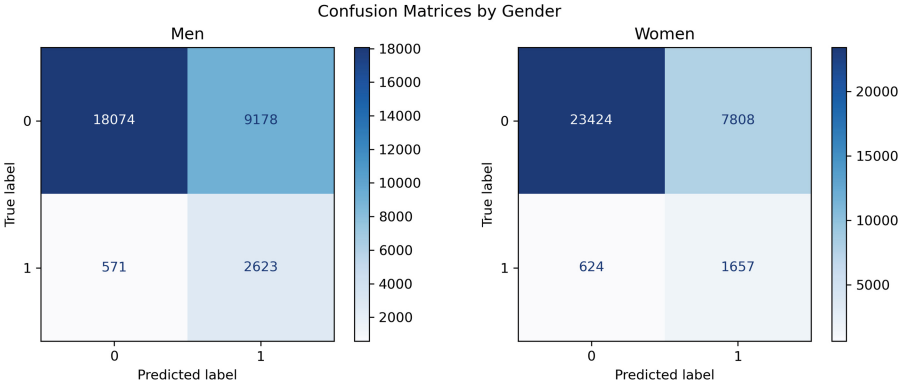


Fig. 7. Confusion matrix stratified by gender (grid search).

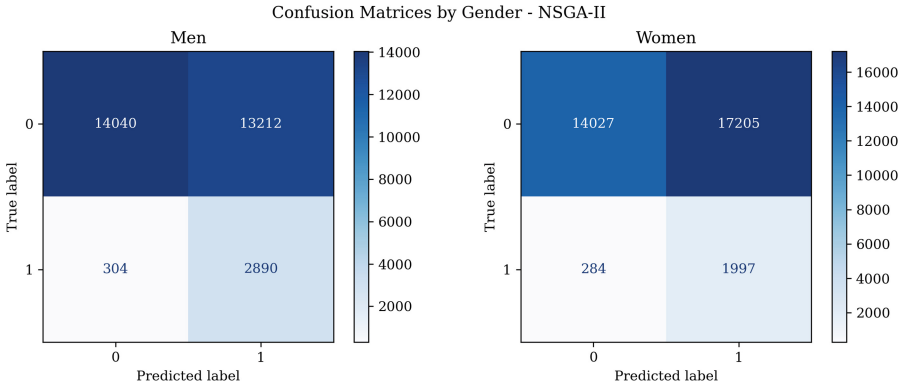


Fig. 8. Confusion matrix stratified by gender (NSGA-II).

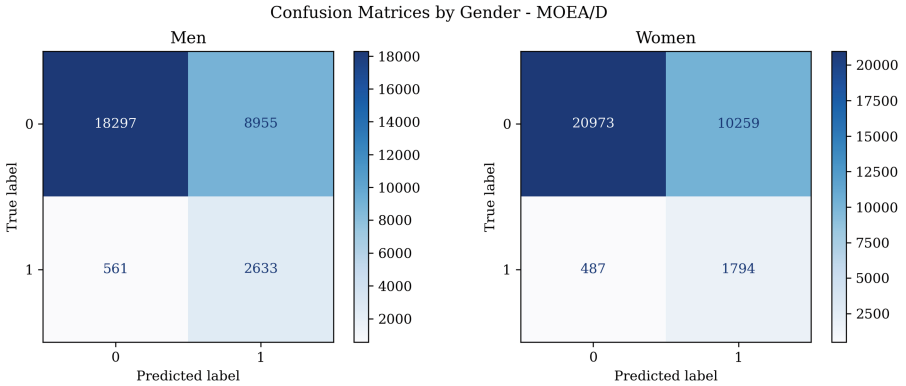


Fig. 9. Confusion matrix stratified by gender (MOEA/D).

strongest discriminative performance, but presents higher sensitivity for men than for women. NSGA-II is the most computationally efficient, but it produces more extreme classification patterns. MOEA/D achieves competitive predictive performance with more balanced error distribution across demographic groups.

4 Conclusions and Future Work

This study investigated the use of multi-objective hyperparameter optimisation to balance predictive performance and fairness in decision tree models. The optimisation problem was formulated with two objectives: minimising classification error and gender-based unfairness, which is measured as disparities in true positive rates under the Equal Opportunity criterion. To analyse the resulting trade-offs, an exhaustive grid search was compared with two evolutionary multi-objective algorithms, NSGA-II and MOEA/D.

The results show that grid search identified the best configuration overall according to the AUC value. However, this exhaustive approach required the greatest computational effort, highlighting its limitations as the hyperparameter space expands. Among the evolutionary approaches, NSGA-II proved to be the most computationally efficient, significantly reducing runtime while still identifying competitive solutions. Nevertheless, its best configurations demonstrated lower discriminative performance (i.e., lower AUC values) compared to those obtained through grid search. In contrast, MOEA/D was able to discover configurations with predictive performance very close to the grid-search optimum, demonstrating its ability to approximate high-quality regions of the Pareto front. Although its runtime was comparable to that of grid search, the resulting models exhibited more balanced false-positive rates across genders, suggesting favourable behaviour in terms of fairness.

The analysis of gender-stratified confusion matrices further revealed that aggregate fairness metrics may obscure important group-level patterns. Even when global unfairness scores appeared small, disparities in recall and false-positive rates remained noticeable. These findings highlight the importance of complementing aggregate fairness measures with detailed group-specific error analysis.

The contribution of this study is to propose a structured multi-objective evaluation framework for fairness-aware hyperparameter tuning. In particular, the comparative analysis of grid search, NSGA-II and MOEA/D highlights their distinct behaviours in terms of fairness–performance trade-offs. The results suggest that evolutionary multi-objective optimisation constitutes a promising alternative to exhaustive search for fairness-aware hyperparameter tuning, especially when computational constraints or large search spaces are present.

This study focuses on decision tree models and a single healthcare dataset. Although being a limitation, this choice allowed for controlled analysis. Future work will extend the analysis to other model families (e.g., ensemble methods and neural networks) with more complex hyperparameter spaces, as well as additional datasets, to assess the robustness and generalisability of the observed patterns.

Fairness was evaluated using the equal opportunity criterion (TPR difference), but other definitions, such as demographic parity and equalised odds, capture different ethical trade-offs and may lead to different model selections. Their incorporation into the multi-objective framework remains an important direction for future research.

Acknowledgements. This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the project grants UIDB/05037/2025 (DOI: 10.54499/UID/05037/2025) and UID/00308/2025 (DOI: 10.54499/UID/00308/2025).

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

References

1. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012). https://dl.acm.org/doi/10.5555/2188385.2188395
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Blank, J., Deb, K.: Pymoo: multi-objective optimization in Python. IEEE Access **8**, 89497–89509 (2020). https://doi.org/10.1109/ACCESS.2020.2990567
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont (1984)
5. Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness. J. Mach. Learn. Res. **24**(312), 1–117 (2023). http://jmlr.org/papers/v24/22-1511.html
6. Cruz, A., Lawrynowicz, A., Bizarro, P.: Promoting fairness through hyperparameter optimization. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 114–123. IEEE (2021)
7. d'Alosio, G., Hort, M., Moussa, R., Sarro, F.: FairRF: multi-objective search for single and intersectional software fairness. In: Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-SEIS '26), Rio de Janeiro, Brazil, 12–18 April 2026. ACM, New York (2026)
8. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms. Wiley, Chichester (2001)
9. 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). https://doi.org/10.1109/4235.996017
10. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
11. Feurer, M., Hutter, F.: Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds.) Automated Machine Learning. TSSCML, pp. 3–33. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05318-5_1
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NIPS 2016), vol. 29, pp. 3315–3323. Curran Associates, Red Hook (2016). https://doi.org/10.48550/arXiv.1610.02413

13. The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 115:1–115:35 (2021). https://doi.org/10.1145/3457607
15. Morales-Hernández, A., Van Nieuwenhuyse, I., Rojas Gonzalez, S.: A survey on multi-objective hyperparameter optimization algorithms for Machine Learning. arXiv preprint arXiv:2111.13755 (2021)
16. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
17. Uddin, S., Lu, H., Rahman, A., Gao, J.: A novel approach for assessing fairness in deployed machine learning algorithms. Sci. Rep. **14**, 17753 (2024). https://doi.org/10.1038/s41598-024-68651-w
18. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comput. **11**(6), 712–731 (2007). https://doi.org/10.1109/TEVC.2007.892759

A Method for Automating Iterative Machine Learning Model Retraining on Event Streams in Decision Support Systems

Arseniy Lomakin[1,2(✉)] and Aleksandr Makarov[1]

[1] Volgograd State Technical University, Lenin avenue, 28, Volgograd 400005, Russia
arseny.lomakin@gmail.com, a.o.makarov2003@gmail.com
[2] Volgograd State Medical University, Fallen Fighters Square, 1, Volgograd 400066, Russia

Abstract. Clinical Decision Support Systems (CDSS) increasingly rely on machine learning models that suffer performance degradation under concept drift in streaming clinical data. Existing approaches typically address drift detection, anomaly identification, or continual learning in isolation, lacking a unified framework with formal non-degradation safeguard. This paper proposes an event-driven continual learning method (Method ECLA) for automated lifecycle management of classifier ensembles in CDSS. The approach integrates Mahalanobis distance-based anomaly detection, clustering-driven micro-batch accumulation, and stability-preserving local adaptation via Elastic Weight Consolidation and Experience Replay. A formal non-degradation constraint, enforced via an automatic rollback mechanism, ensures that model updates do not reduce predictive performance beyond a predefined tolerance ($\varepsilon = 0.05$). The method is evaluated on 70,421 real ophthalmological cases over a 180-day operational horizon across 10 independent runs. Experimental results demonstrate that Method ECLA (Event-driven Continual Learning with Anomaly clustering) outperforms static and periodically retrained baselines, achieving Accuracy of 0.890 and AUC-ROC of 0.873, while reducing Time-to-Recovery (TTR) to 1.2 days and maintaining computational efficiency independent of dataset size. The proposed Method ECLA enables stable, adaptive, and resource-efficient deployment of machine learning models in real-world CDSS environments.

Keywords: Clinical Decision Support Systems · Continual Learning · Concept Drift · Anomaly Detection · Ensemble Learning · IOL Power Calculation

1 Introduction

Machine learning models are increasingly being integrated into Clinical Decision Support Systems (CDSS) to help physicians make more accurate diagnoses and select the appropriate treatment plan. Such systems use large volumes of clinical data to train models that assist in medical decision-making.

However, machine learning models implemented in Clinical Decision Support Systems (CDSS) in real-world medical practice are inevitably subject to obsolescence.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 292–308, 2027.
https://doi.org/10.1007/978-3-032-30494-0_19

Advances in diagnostic methods in modern medicine lead to the identification of previously unknown diseases and the refinement of classifications of existing nosological forms, which allows for the development of more precise treatment plans for each individual patient. This gives rise to the phenomenon of conceptual drift [1–3], in which the statistical relationship between input features (e.g., patient condition data) and target variables (e.g., the proposed diagnosis) changes, which over time leads to a decrease in the accuracy and predictive value of machine learning models. As a result, the likelihood of incorrect system recommendations increases, which potentially raises the risk of medical errors and, consequently, jeopardizes the quality of patient care.

Traditional approaches to maintaining the quality of machine learning models are based on periodic retraining. Scheduled retraining involves updating models on new datasets according to a predefined schedule, including instances of full retraining using the entire available dataset. Despite its relative simplicity of implementation, this practice cannot ensure effective adaptation to the actual flow of clinical data. During periods of stability in the data distribution, computational resources are wasted on scheduled retraining cycles. However, in situations of sudden shifts in the distribution of input data, the model remains outdated until the scheduled retraining time arrives, and in the case of full retraining, excessive computational costs and a decrease in accuracy on outliers are observed.

Traditional approaches include hyperparameter optimization using GridSearch. This method allows determining the optimal model configuration on a fixed dataset; however, it does not provide mechanisms for responding to outliers and does not include a trigger threshold for initiating retraining upon detection of conceptual drift, thereby requiring manual retriggering.

Another traditional approach is the incremental neural network (INN) approach with step-by-step training. Despite its ability to gradually learn patterns, it exhibits low accuracy and requires manual retraining whenever new data is accumulated.

One alternative approach is the online updating of model weights, which ensures adaptation to the changing distribution of clinical data as each new example is received. However, such retraining carries the risk of catastrophic forgetting—a phenomenon in which the model loses previously learned patterns when trained on new data [4].

The authors set out to automate the iterative training of machine learning models in Clinical Decision Support Systems (CDSS) based on an event stream by applying the developed Method ECLA, which reduces the resource costs of retraining as the data volume increases while adhering to three key criteria: a safeguard of no degradation; reactivity—response upon the occurrence of an anomaly; and adaptability—dynamic expansion of recognizable patterns.

This article discusses the stages of creating and implementing the proposed Method ECLA, evaluates its key properties, conducts a comparative analysis with existing approaches, and presents a software tool developed for this method.

The contributions of this work are as follows:

- An event-driven continual learning framework for CDSS that unifies anomaly detection, micro-batch accumulation, and stability-preserving local adaptation under a formal non-degradation safeguard;

- A statistically principled anomaly clustering mechanism that forms representative micro-batches without dependence on the total dataset size, ensuring constant-cost retraining;
- A dual-memory adaptation strategy combining Elastic Weight Consolidation and Experience Replay, which prevents catastrophic forgetting while learning new clinical patterns;
- Comprehensive experimental validation on 70,421 ophthalmological cases, demonstrating stable performance over 180 days and significant improvements over currently deployed retraining strategies.

2 Methodology

2.1 Proposed Solution

In existing research on the application of machine learning in CDSS, adaptation to conceptual drift, anomaly detection, and continuous learning are typically addressed in isolation [5–7]. However, in real-world operating conditions, CDSS function in non-stationary environments where these problems may arise simultaneously [8, 9]. Consequently, effective long-term implementation requires a unified approach that ensures: 1) timely detection of shifts in data distribution; 2) stable and controlled model updates; 3) resistance to catastrophic forgetting of previously learned patterns; 4) a safeguard that performance remains within a specified threshold during model adaptation.

Based on the literature review, it was found that existing approaches often fulfill only a few of these requirements [10–12]. The absence of a unified formalized structure that ensures the simultaneous fulfillment of all four requirements justified the need to develop and experimentally validate the proposed method ECLA.

In practice, clinical data in CDSS arrive sequentially and can be formalized as a stream of paired observations (1):

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\} \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the feature vector of the i-th clinical case, and $y_i \in \mathbb{R}$ denotes the corresponding target variable. This formulation reflects the continuous accumulation of clinical knowledge and requires incremental, controlled model updates under conditions of potential distribution shifts.

Let the ensemble of machine learning models deployed in the CDSS at time t be defined as (2):

$$M_t = \left\{M_t^1, M_t^2, \ldots, M_t^n\right\} \tag{2}$$

where each model $M_t^i (i = 1, 2, \ldots, n)$ performs a classification task aimed at selecting the most appropriate formula (Barrett, Haigis, Hoffer Q, Holladay) for calculating the power of an intraocular lens (IOL) for a given clinical case.

The ensemble's performance is evaluated using the quality function $Q(M_t)$, which reflects the predictive ability of the entire ensemble of models on a fixed dataset at time t. In practical applications, this function can be interpreted as standard classification quality metrics or AUC-ROC.

From the perspective of the machine learning lifecycle [13], CDSSs are dynamic systems that require constant maintenance of the relevance of predictive models. The main task lies not in initial training, but in controlled adaptation to new clinical data over time. In this regard, an appropriate mechanism for updating machine learning models is required, one that is not prone to degradation and instability during periods of changing data distribution.

Method ECLA formalizes a constraint on the preservation of the model's predictive performance, defined as (3)

$$Q(M_{t+1}) \geq Q(M_t) - \varepsilon, \tag{3}$$

where $Q(M_t)$ is a scalar measure of model quality at iteration t, calculated based on standard classification metrics (Accuracy, AUC-ROC, F1-score), and ε ($\varepsilon = 0.05$) is a predefined tolerance parameter, empirically selected to balance model adaptability with performance stability during data updates. In this study, Accuracy and AUC-ROC are used as the primary metrics for evaluating predictive capabilities, which is consistent with common practice in evaluating machine learning models in medical applications [14].

Thus, the goal of the proposed method is to automatically manage the lifecycle of machine learning models in a clinical data stream while simultaneously meeting key requirements: responsiveness, adaptability, and quality preservation. To achieve this goal, an event-driven method for iterative ensemble learning of classifiers has been developed to support clinical decision-making when selecting the optimal formula for calculating the optical power of an IOL replacing a clouded lens.

It should be noted that the proposed method is not limited solely to cataract surgery. The basic framework—event-driven clustering of anomalies and local fine-tuning with a non-degradation safeguard—is applicable to any CDSS in which an ensemble of classifiers operates on a data stream subject to conceptual drift. The scope of application can be extended to other surgical specialties, diagnostic tasks, and medical information systems.

2.2 Anomaly Detection Mechanism

One of the key components of Method ECLA is the identification of clinical observations that deviate from the statistical distribution of the training dataset. The occurrence of such observations may indicate the emergence of new patterns, so the CDSS must respond adaptively to such cases.

To detect novelty in the incoming stream of clinical data, Method ECLA uses a metric such as the Mahalanobis distance, which measures the statistical deviation of a new observation from the multidimensional distribution of the training dataset [15].

A new clinical case, represented by the feature vector x, is considered an anomalous event e_a if condition (4) is satisfied:

$$D_m(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} > \tau \tag{4}$$

where μ is the sample mean of the feature vector of the training sample, Σ is the covariance matrix of the feature space, and τ is the novelty detection threshold.

The choice of the Mahalanobis distance is based on the statistical properties of the ophthalmic parameters used in IOL calculations: gender (GENDER) and the eye being examined (EYE); biometric parameters of the eye—axial length (AL), anterior chamber depth (ACD), lens thickness (LT), white-to-white distance (WTW); keratometric parameters of the anterior surface (K1, K2, ΔK), total (TK1, TK2, ΔTK), and posterior surface (PK1, PK2, ΔPK2); corneal thickness (CCT), Kappa angle (CW-Chord), and refractometry (AxR); calculated values of the spherical equivalent and optical power of the IOL using four formulas: Barrett, Haigis, Hoffer Q, Holladay. The listed characteristics are mutually correlated and measured on different scales and in different units of measurement.

Under such conditions, the Euclidean distance becomes statistically incorrect, as it assumes the independence and equal-scalability of the features. However, the Mahalanobis distance explicitly accounts for the covariance structure via the matrix Σ^{-1}, which makes it invariant to the scales and correlations of the features.

The novelty detection threshold τ is derived mathematically. Assuming that the training data follow a multivariate normal distribution—which is approximated by an empirical estimate of the covariance matrix—the square of the Mahalanobis distance follows a chi-squared distribution with k degrees of freedom (5):

$$D_m^2(x) \sim \chi_k^2 \tag{5}$$

where k denotes the dimension of the feature space.

Consequently, the threshold value τ is chosen as (6):

$$\tau = \sqrt{\chi^2_{k,1-a}} \tag{6}$$

where α denotes the significance level, which determines the probability of a false positive detection of an anomaly in the input data.

In this study, the feature space has a dimension of $k = 33$, corresponding to 32 biometric parameters measured by ophthalmic devices (IOL Master 700, Tomey OA-2000, Canon RK-1, Sirius, Pentacam). At a significance level of $\alpha = 0.01$, the obtained threshold is $\tau \approx 4.82$. This means that no more than 1% of standard clinical observations will be incorrectly classified as anomalies, thus maintaining the false positive rate below 1%.

It is worth noting that the detected anomalous data are not used immediately to retrain the model. Instead, they are accumulated and subjected to further processing using the clustering mechanism described in the following subsection.

2.3 Anomaly Clustering and Micro-batch Formation

The clustering scheme for newly detected patterns, implemented in Method ECLA, helps avoid instability caused by rare isolated anomalies that may occur in incoming streams of medical data. In other words, this approach allows the system to accumulate a statistically significant number of anomalous observations before beginning model adaptation.

Each anomalous observation x_i is routed to the nearest cluster of anomalies according to the Mahalanobis distance. The optimal cluster is defined as (7):

$$C_{opt} = \text{argmin}_k D_m(x_i, \text{centroid}(C_k)) \tag{7}$$

where $\text{centroid}(C_k)$ denotes the centroid of the kth cluster of anomalies.

To classify incoming medical data, the Mahalanobis distance from the current observation x_i to the center of the nearest cluster C_{opt} is calculated. If the distance between x_i and C_{opt} satisfies condition (8):

$$D_m(x_i, \text{centroid}(C_{opt})) \leq \gamma \tag{8}$$

where γ denotes the decision threshold that distinguishes the region of acceptable statistical deviations in the data from the region of novelty; in this case, the method interprets the observation as belonging to a previously detected anomaly type, and it is added to the corresponding cluster. Otherwise, this indicates the emergence of a previously undetected clinical pattern, and a new cluster of anomalies is created (9).

$$C = C \cup \{\{x_i\}\} \tag{9}$$

Model retraining does not start immediately after a new cluster is formed. Instead, the system accumulates outliers until the cluster itself reaches a predefined minimum size.

Method ECLA includes the local retraining procedure LearnLocalExpert(C_{opt}), which is triggered if condition (10) is satisfied:

$$|C_{opt}| \geq N_{th} \tag{10}$$

where N_{th} is the threshold value for the minimum cluster size.

The lower bound of the micro-batch threshold $N_{th} = 33$ is justified by the condition of non-degeneracy of the covariance matrix. In a d-dimensional feature space, at least $d + 1$ observations are required for a stable estimation of the covariance matrix. With 32 parameters measured by ophthalmic devices, $d = 32$, from which it follows that the minimum required number of observations is 33. In turn, the upper bound of the micro-batch threshold $N_{th} = 50$ was determined based on an analysis of continuous learning metrics. The theoretical dependence of the Forgetting Measure (FM) and Forward Transfer (FWT) [16] on the batch size N is shown in Fig. 1.

Thus, this interval $N_{th} \in [33, 50]$ represents a scientifically validated optimal range that simultaneously minimizes the FM index and maximizes the FWT index, thereby ensuring a balance between the stability of previously learned patterns and the model's ability to adapt.

2.4 Continual Learning and Catastrophic Forgetting Prevention

The local retraining procedure LearnLocalExpert(C_{opt}), which was introduced and described in the previous subsection, should take into account newly discovered clinical patterns without compromising the model's predictive performance for previously

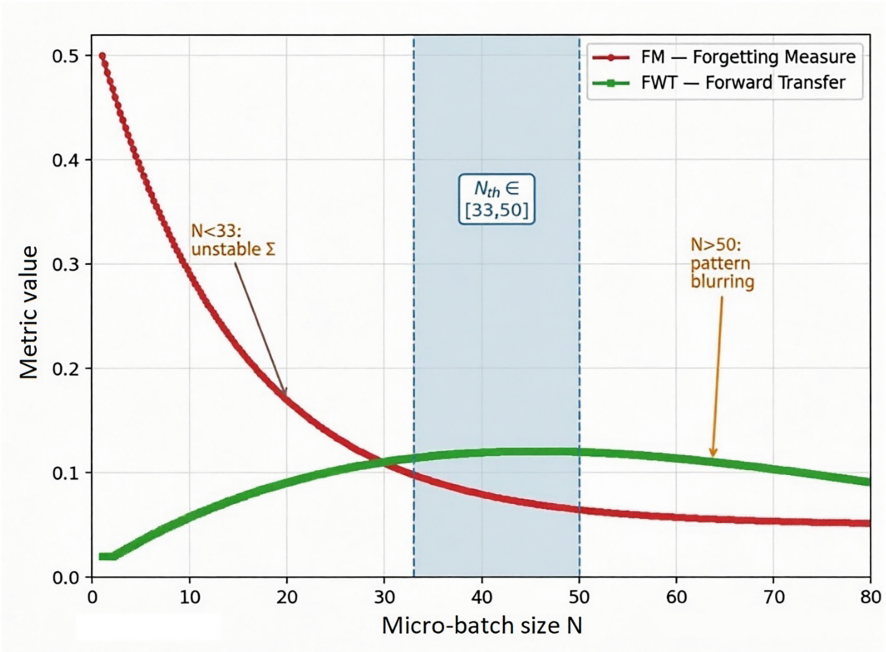


Fig. 1. The dependence of the Forgetting Measure (FM) and Forward Transfer (FWT) on the size of the micro-batch N.

recorded observations. It should be noted that this procedure does not involve creating a new model, but rather retraining the current ensemble of classifiers M_t by fine-tuning its weights. In continuous learning, simple retraining on newly acquired data can lead to catastrophic forgetting, resulting in the loss of previously learned patterns when adapting to new cases. To avoid this, in Method ECLA this problem is solved using two complementary mechanisms, namely Elastic Weight Consolidation (EWC) [17] and Experience Replay [18].

The EWC approach adds a regularization term to the standard loss function, "locking in" the values of weights that are critically important for historical data. The importance of each weight, in turn, is estimated via the diagonal elements of the Fisher matrix, which allows for differentiated penalization of parameter changes. The resulting loss function takes the form (11):

$$L_{EWC}(\theta) = L_B(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2 \quad (11)$$

where $L_B(\theta)$ denotes the loss function calculated on the current data batch, θ_i^* represents the optimal model weights obtained on the training dataset D, F_i are the diagonal elements of the Fisher matrix, which quantitatively reflect the importance of each weight parameter for preserving predictive power on previously studied distributed data, and λ is the regularization coefficient that controls the balance between learning on new data and preserving historical knowledge. Thus, EWC selectively constrains important weights, allowing plasticity in less critical parameters.

The Fisher information matrix is estimated as the mathematical expectation of the square of the gradient of the log-likelihood with respect to the model parameters (12). This approach quantitatively measures the sensitivity of the likelihood function to changes in each weight. Elements with a high value of F_i correspond to parameters that strongly influence the quality of the input data—changes to them during fine-tuning are penalized more heavily.

$$F_i = E\left[\left(\frac{\partial \log p(D|\theta)}{\partial \theta_i}\right)^2\right] \tag{12}$$

In the proposed system, the EWC mechanism is based on an optimized C++ implementation integrated into the PyTorch framework; since the Fisher matrix is computed based on the log-likelihood gradients for all historical data, its full evaluation requires significant computational resources in a naive implementation.

The second mechanism used in the retraining procedure is Experience Replay, which forms a training batch as a combination of a new anomaly cluster C_{opt} and a random sample from the historical data buffer $\text{Buffer}(D, n_{replay})$ (13):

$$B_{train} = C_{opt} \cup \text{Buffer}(D, n_{replay}) \tag{13}$$

The fact that historical examples are present in every fine-tuning batch ensures that the model continues to analyze standard clinical cases while specializing in the new pattern.

Thus, with naive fine-tuning on only C_{opt}, the accuracy on historical data degrades to Accuracy $\approx$ 0.20. In turn, the method described in this subsection maintains the quality of historical data above the threshold value $\varepsilon = 0.05$, thereby satisfying the non-degradation criterion.

2.5 Event-Driven Iterative Learning Workflow

The mechanisms described above constitute the key components of Method ECLA. This subsection presents a formalized workflow that governs the model's lifecycle in the context of clinical data stream processing. Algorithm 1 shows the proposed method in the formal form.

Algorithm 1 Event-driven iterative learning (Method ECLA)

Input: Initial ensemble M_t, clinical data stream $D = \{(x_i, y_i)\}$, novelty detection threshold τ, cluster acceptance threshold γ, micro-batch size $N_{th} \in [33, 50]$, non-degradation threshold ε
Output: Updated M ensemble

```
Initialize anomaly cluster set C ← ∅
Initialize current model M ← M_t
for each incoming observation x_i ∈ D do
    // --- Event e_a: Anomaly detection ---
    Compute D_m(x_i)
    if D_m(x_i) > τ then
        e ← e_a
        // --- Clustering and accumulation ---
        Find C_opt = argmin_k D_m(x_i, centroid(C_k))
        if D_m(x_i, centroid(C_opt)) ≤ γ then
            C_opt ← C_opt ∪ {x_i}
        else
            Create new cluster C_new = {x_i}
            C ← C ∪ {C_new}
            C_opt ← C_new
        end if
        // --- Event e_th: Micro-batch accumulation ---
        if |C_opt| ≥ N_th then
            M' ← LearnLocalExpert(M, C_opt)
            // --- Non-degradation check ---
            if Q(M') ≥ Q(M) − ε then
                M ← M'
            else
                M ← Rollback(M)
            end if
        end if
    else
        // --- Event e_t: Time-based retraining and Rollback ---
        if TimeIntervalExceeded() then
            M' ← GlobalRetrain(M, D)
            if Q(M') ≥ Q(M) − ε then
```

The operation of Method ECLA is determined by a continuous stream of medical data events (14) that enter the database after processing the biometric parameters measured by ophthalmic devices.

$$E = \{e_1, e_2, \ldots, e_n\} \tag{14}$$

In the context of the proposed solution, three types of events were defined: anomaly event e_a—detection of a clinical case that statistically deviates from the distribution of data learned during the training phase; the data accumulation event e_{th}—the anomaly cluster reaching the minimum micro-batch size required for retraining; the time interval event e_t—the expiration of a specified time interval since the last update.

The developed method formalizes a hierarchy of event priorities: e_a, then e_{th}, then e_t, which is justified by an empirical analysis of continuous learning metrics.

An anomaly event has the highest priority, as it signals the emergence of a previously unobserved clinical pattern requiring an immediate system response. In this case, the newly observed data is processed by the anomaly clustering mechanism, described in detail in Sect. 2.3.

When the size of the anomaly cluster reaches a set minimum threshold, a data accumulation event occurs. This triggers the local model adaptation procedure LearnLocalExpert(C_{opt}), which is described in Sect. 2.4—this is the system's primary adaptation mechanism.

The time interval event serves as a safety mechanism for Method ECLA. If, within a specified time, no anomaly cluster has reached the minimum N_{th} threshold, the system performs global retraining on the entire dataset currently available.

This hierarchy ensures the optimal use of computational resources: computationally intensive global retraining is performed only, when necessary, while local adaptations to new patterns are initiated as data accumulates.

After each retraining procedure, the updated ensemble M'_{t+1} is evaluated using a quality function. The system's transition from state S_t to state S'_t is permitted only if condition (15) is satisfied:

$$Q\left(M'_{t+1}\right) \geq Q(M_t) - \varepsilon \tag{15}$$

This approach ensures that the system's forecasting quality remains within a specified threshold for acceptable degradation.

In addition to the quality degradation limit, Method ECLA implements an automatic Rollback(M_t). If the updated ensemble does not meet the quality criterion, the system automatically restores the previous state of the model, preventing the deployment of versions with degraded predictive performance.

The complete algorithm for Method ECLA—from event registration in the data stream to updating the classifier ensemble or performing a rollback when the non-degradation criterion is violated—is shown in Fig. 2.

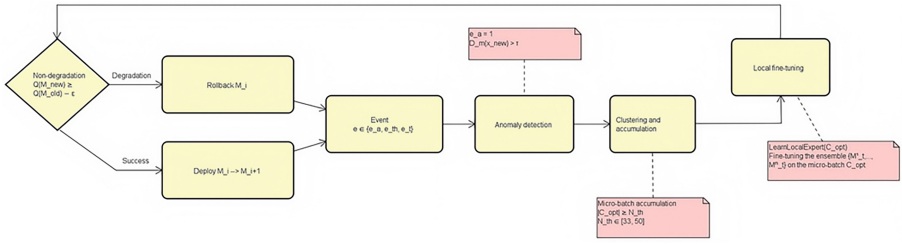


Fig. 2. Flowchart of Method ECLA for automatic iterative learning.

The diagram illustrates a three-level hierarchy for event processing, an anomaly clustering mechanism, quality verification of the updated model, and an automatic rollback procedure.

Thus, the architecture of the developed Method ECLA ensures stable and resource-efficient continuous learning in CDSS under conditions of non-stationary clinical data distributions.

3 Results

3.1 Experimental Setup

The authors conducted an experimental validation of Method ECLA to assess its effectiveness in a CDSS environment. The study utilized an ensemble of classifiers from the "LensCalc" system, designed to assist ophthalmologists in calculating IOLs based on biometric parameters.

The experiment utilized a dataset comprising $|D| = 70\,421$ clinical cases (MNTK, 2018–2025), divided using a stratified method into a training set D_{train} (80%) and a test set D_{test} (20%) subsets. The test set D_{test} remained unchanged across all experiments to ensure the comparability of results.

The extraction of clinical data was automated using an OCR parser (PyTesseract) to process diagnostic reports in PDF format. The raw data, stored in sixth normal form (6NF), was converted to third normal form (3NF) by consolidating one record per patient and aggregating measurements from multiple devices. For duplicate measurements, priority was given to values from devices with higher accuracy; otherwise, the most recent measurement was retained. The input data for the ensemble of classifiers includes 32 clinical parameters described in Sect. 2.2.

Each classifier in the ensemble generates a binary prediction indicating whether a specific IOL calculation formula is applicable (1) or not (0). Ground-truth labels were determined retrospectively: a formula was considered applicable if the postoperative refractive error deviated from the target by no more than ± 0.50 diopters.

To assess the stability of the method, $n = 10$ independent runs were performed with random initial ensemble states (initial accuracy $Acc \in [0.77, 0.83]$) over a horizon of 100 iterations, corresponding to 180 days of system operation.

The initial states for all experiments are shown in Table 1.

Table 1. Initial conditions M_0 for 10 independent experiments.

Parameter	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
Accuracy M_0	0.79	0.83	0.81	0.77	0.82	0.80	0.78	0.83	0.79	0.81
AUC M_0	0.80	0.84	0.82	0.78	0.83	0.81	0.79	0.84	0.80	0.82

This experimental setup ensures the long-term stability of Method ECLA under conditions of gradual data accumulation and shifts in the distribution.

3.2 Basic Retraining Methods for Comparison

To evaluate the effectiveness of the proposed method, we compared it with three baseline retraining methods widely used in practical machine learning systems.

The first (B_1) is periodic manual retraining at 14-day intervals. This approach reflects the traditional CDSS workflow, where updates are performed infrequently due to high computational costs and the need for manual validation.

The second baseline method (B_2) is automatic scheduled retraining at 7-day intervals without data distribution analysis, using the Cron job scheduler. This approach reduces the need for manual intervention but does not adapt to conceptual data drift.

The third baseline method (B_3) is automatic retraining using GridSearch every 7 days. This approach ensures full adaptation of the model's hyperparameters based on new data, but requires significant computational resources due to the exhaustive search of all combinations.

All methods were evaluated under the identical experimental conditions described in the previous subsection, ensuring the comparability of results, and differences in performance are determined solely by the retraining strategy.

These baselines reflect prevailing maintenance strategies in medical information systems; advanced streaming and continual learning extensions are addressed in the limitations.

3.3 Comparison Results

A comparative analysis of the developed Method ECLA and the baseline retraining methods described in Sect. 3.2 was conducted in accordance with the experimental setup defined earlier in Sect. 3.1.

The evaluation included both measures of predictive performance and measures of the method's stability over time. Prediction quality was assessed using the Accuracy and AUC-ROC metrics, while performance stability during long-term operation was evaluated using the FM and Time-to-Recovery (TTR) metrics.

The Forgetting Measure quantitatively assesses the level of performance degradation accumulated during iterative model updates and is defined as follows (16):

$$FM = \max_{t<T}[Q(M_t) - Q(M_T)] \tag{16}$$

where $Q(M_t)$ represents the value of the quality metric at iteration t, and T corresponds to the final iteration of the experiment.

The TTR metric characterizes the time interval required for the system to recover its predictive capability after a degradation event and is calculated using the following formula (17):

$$TTR = t_{recovery} - t_{event} \tag{17}$$

where t_{event} denotes the specific time at which a performance degradation is observed, and $t_{recovery}$ corresponds to the moment when the model's performance returns to its pre-t_{event} state.

To assess the statistical significance of the identified differences in the predictive ability of the methods, the nonparametric Mann-Whitney U-test [19] was applied for independent samples (sample size $n = 10$). The formulation of the hypotheses to be tested is given by (18):

$$H_0: Q(M_t) = Q(M_0), H_1: Q(M_t) > Q(M_0) \tag{18}$$

at a significance level of $\alpha = 0.05$. The summary results of the experiment are presented in Table 2.

Table 2. Results of a comparative experiment (180-day time horizon, n = 10 independent runs).

Criterion	Manual retraining	Cron (7d)	GridSearch (7d)	Method ECLA
Accuracy (t = 100)	0.800	0.830	0.835	0.890
AUC-ROC (t = 100)	0.810	0.840	0.830	0.873
Median TTR, d	14.0	7.0	7.0	1.2
FM (medium)	0.081	0.062	0.058	0.019
No-degradation safeguard	No	No	No	Yes (rollback)
Drift adaptability	No	No	Partial	Yes
Resources/cycle, h	25	25	25+	2–12

The trend in the AUC metric over the course of 10 experimental iterations (Fig. 3) indicates that the proposed Method ECLA exhibits a steady, monotonically increasing performance, whereas the baseline methods reach a plateau as early as iterations (t) 3 and 4.

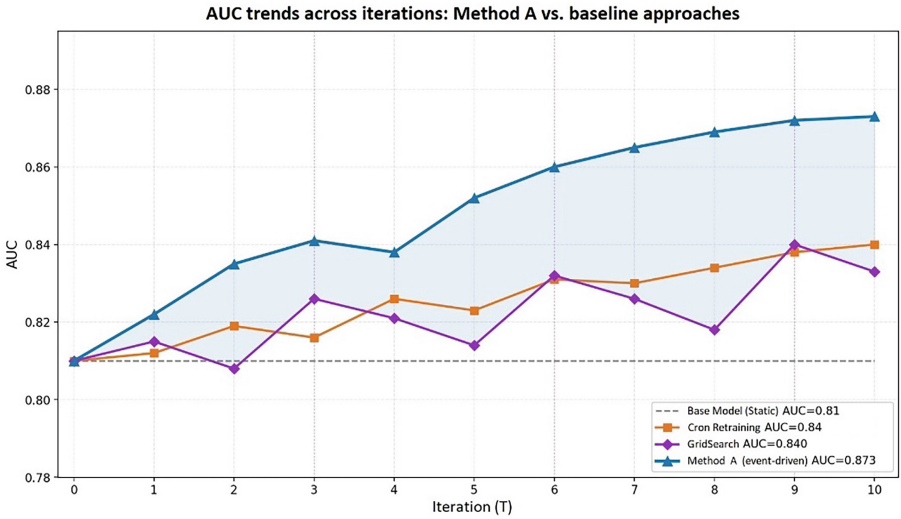


Fig. 3. Changes in the AUC-ROC curve over the course of the iterations.

A comparative analysis of the methods based on classification accuracy and time-to-response (TTR) is presented in Fig. 4.

The statistical significance of the increase in accuracy was verified using the Mann-Whitney U-test. In the early stages of the experiment (at t = 1 and t = 2), the differences in performance were not statistically significant (p = 0.485 and p = 0.285

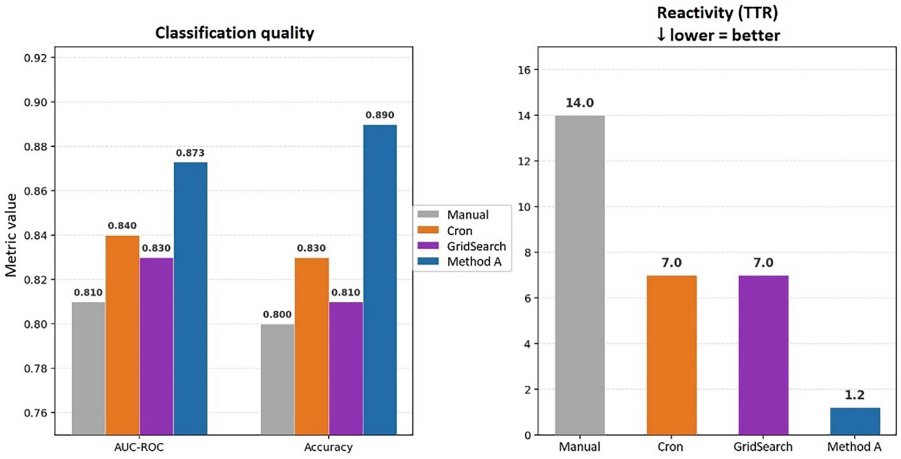


Fig. 4. Comparison of four methods based on AUC-ROC, accuracy, and median TTR in days.

respectively), which is explained by the fact that the volume of accumulated data had not yet reached the threshold value N_{th}. Starting from iteration $t = 5$, the first statistically significant increase in performance was recorded ($p = 0.032$). At $t \geq 10$, a significant difference is observed ($p < 0.001$, $U = 100$), which confirms the complete superiority of Method ECLA: each of the 10 final states of the M_t model demonstrates higher quality metrics compared to all 10 initial states of M_0. The median value of the Accuracy metric shows a monotonic increase with an increment of $\Delta = +0.113$ (in the interval from $t = 1$ to $t = 100$), accompanied by a decrease in variance, indicating increased stability.

The learning curve for Method ECLA, as well as the trend in the AUC-ROC metric for all four methods under consideration (A, B_1, B_2, B_3) over the 180-day experimental period, are shown in Fig. 5.

The learning curve for Method ECLA, plotted based on the results of 10 independent experimental runs, demonstrates steady convergence toward a stable solution. The convergence gap between training accuracy and test accuracy is 0.027 with 20,000 training instances, which is an indication of the absence of overfitting. The $\pm\sigma$ confidence intervals, calculated based on 10 independent experimental runs, show a consistent narrowing as the sample size increases, indicating improved stability and reproducibility of Method ECLA's results as data accumulates.

Another important result of this work is the confirmation of the scalability of Method ECLA. As the data volume increases from 200 to more than 20,000 clinical cases, the local retraining time remains virtually unchanged, since it is determined by the size of the anomaly cluster ($N_{th} \in [33, 50]$) rather than the total database size. This fundamentally distinguishes Method ECLA from basic approaches, for which global retraining time increases linearly with the volume of data: when scaling from 2,000 to 70,000 cases, the duration of manual retraining was 25 h.

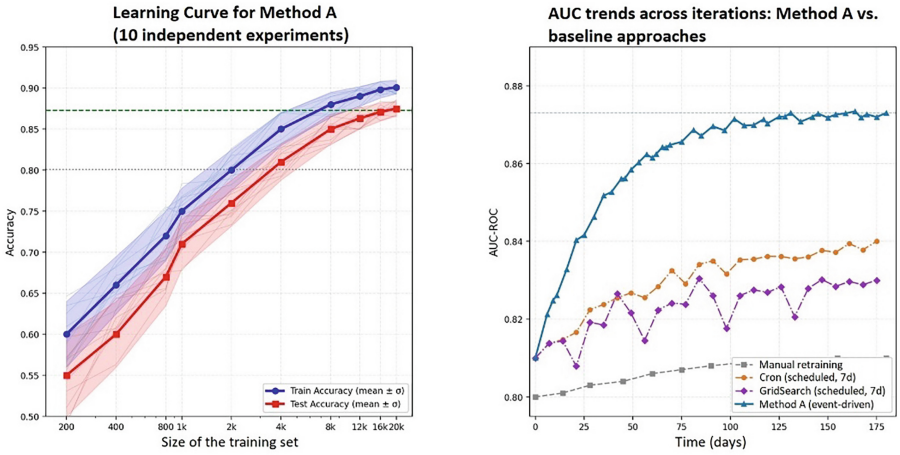


Fig. 5. Learning curve of Method ECLA across 10 independent runs (left panel) and AUC-ROC dynamics for four methods over a 180-day period (right panel).

4 Discussion

The study's findings confirm that the reliable long-term operation of machine learning models in clinical decision support systems requires not only high predictive performance but also robust mechanisms that ensure continuous adaptation to changing data distributions and protection against performance degradation. Traditional retraining strategies typically address only certain aspects of this problem and therefore cannot simultaneously safeguard the model's responsiveness, adaptability, and performance stability.

An important finding of the study is that the superiority of Method ECLA over baseline approaches lies not only in achieving higher final values of the Accuracy and AUC-ROC metrics, as experiments have shown, but also in the stability of performance during the training process.

A significant methodological contribution is the use of a constraint on preserving predictive ability in conjunction with a rollback mechanism $\text{Rollback}(M_t)$. Unlike approaches that rely on implicit assumptions about the reliability of retraining procedures, the proposed method introduces a formal validation criterion for each model update [20–24]. Thus, model adaptation is formalized as a conditional optimization problem with performance preservation safeguard. This constraint is critically important for CDSS, where temporal degradation in prediction quality can directly increase the risk of medical error, thereby jeopardizing the quality of patient care.

A key component of the proposed approach is an anomaly detection procedure based on the Mahalanobis distance, which is particularly applicable to clinical biometric parameters characterized by mutual correlations and heterogeneous measurement scales. A statistically justified threshold derived from the chi-square distribution provides a rigorous mechanism for identifying new clinical patterns that require model adaptation.

While the anomaly detection module assumes multivariate normality, empirical validation confirmed that transformed parameters (log for axial length, keratometry) adhere

to this distribution (Shapiro–Wilk $p > 0.05$). Future ECLA iterations may integrate non-parametric detectors (e.g., Isolation Forest, autoencoders) to mitigate this limitation in heterogeneous populations.

Experimental results demonstrate that integrating incremental learning mechanisms with stability-preserving approaches, including Elastic Weight Consolidation and Experience Replay, enables the system to incorporate new knowledge while preventing catastrophic forgetting. This balance between adaptability and stability is a prerequisite for the sustainable long-term deployment of machine learning systems in medical practice.

An additional advantage of the proposed approach is computational scalability. Since local retraining is triggered by clusters of anomalies rather than the total dataset size, the computational cost of model updates remains stable despite the growing database. This property significantly reduces computational requirements and minimizes the need for manual intervention during system maintenance.

Promising areas for future research include extending the proposed methodology to additional clinical domains and investigating adaptive mechanisms for dynamically optimizing thresholds in anomaly detection procedures.

While the baseline retraining strategies (manual, Cron-based, GridSearch) employed are foundational rather than state-of-the-art in continual learning and concept drift adaptation, this study prioritized validating the feasibility and stability of the proposed event-driven update logic under realistic clinical streaming conditions. Future work will systematically benchmark this approach against advanced methods (e.g., iCaRL, GEM, drift-aware ensembles) to further establish performance boundaries.

References

1. Hinder, F., Vaquet, V., Hammer, B.: One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: detecting concept drift. Front. Artif. Intell. **7**, 1330257 (2024). https://doi.org/10.3389/frai.2024.1330257
2. Hinder, F., Vaquet, V., Hammer, B.: One or two things we know about concept drift—a survey on monitoring in evolving environments. Part B: locating and explaining concept drift. Front. Artif. Intell. **7**, 1330258 (2024). https://doi.org/10.3389/frai.2024.1330258
3. Garcia, C.M., Abilio, R., Koerich, A.L., Britto, A.S., Barddal, J.P.: Concept drift adaptation in text stream mining settings: a systematic review. ACM Trans. Intell. Syst. Technol. **16**(2), 1–67 (2025). https://doi.org/10.1145/3704922
4. Li, H., Ding, L., Fang, M., Tao, D.: Revisiting catastrophic forgetting in large language model tuning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Findings of ACL: EMNLP 2024, pp. 4297–4308. ACL (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.249
5. Samariya, D., Ma, J., Aryal, S., Zhao, X.: Detection and explanation of anomalies in healthcare data. Health Inf. Sci. Syst. **11**(1), 20 (2023). https://doi.org/10.1007/s13755-023-00221-2
6. McLean, C., Capurro, D.: Concept drift detection to assess the diffusion of process innovations in healthcare. In: AMIA Annual Symposium Proceedings, pp. 746–755 (2023)
7. Kavya, R., et al.: DAAR: drift adaptation and alternatives ranking approach for interpretable clinical decision support systems. Biomed. Signal Process. Control **84**, 104793 (2023). https://doi.org/10.1016/j.bspc.2023.104793
8. Lasko, T.A., Strobl, E.V., Stead, W.W.: Why do probabilistic clinical models fail to transport between sites. npj Digit. Med. **7**(1), 53 (2024). https://doi.org/10.1038/s41746-024-01037-4

9. Chen, Z., et al.: Harnessing the power of clinical decision support systems: challenges and opportunities. Open Heart **10**, e002432 (2023). https://doi.org/10.1136/openhrt-2023-002432
10. da Silva Neto, S.R., et al.: Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: a systematic review. PLoS Negl. Trop. Dis. **16**(1), 1–37 (2022). https://doi.org/10.1371/journal.pntd.0010061
11. Susanto, A.P., Lyell, D., Widyantoro, B., Berkovsky, S., Magrabi, F.: Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J. Am. Med. Inform. Assoc. **30**(12), 2050–2063 (2023). https://doi.org/10.1093/jamia/ocad180
12. Behrendt, F., et al.: Leveraging the mahalanobis distance to enhance unsupervised brain MRI anomaly detection. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 394–404. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_37
13. Xie, Y., Cruz, L., Heck, P., Rellermeyer, J.S.: Systematic mapping study on the machine learning lifecycle. arXiv preprint arXiv:2103.10248 (2021)
14. Collins, G.S., et al.: TRIPOD+AI statement: Updated guidance for reporting clinical prediction models. BMJ **385**, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378
15. Zozoulenko, N., Cass, T., Gonon, L.: Infinite-dimensional Mahalanobis distance with applications to kernelized novelty detection. arXiv preprint arXiv:2407.11873 (2025)
16. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. IEEE Trans. Pattern Anal. Mach. Intell. **46**(8), 5362–5383 (2024). https://doi.org/10.1109/TPAMI.2024.3367329
17. Aslam, S., Rasool, A., Li, X., Wu, H.: CEL: a continual learning model for disease outbreak prediction by leveraging domain adaptation via elastic weight consolidation. Interdiscip. Sci. Comput. Life Sci. **17**(2), 390–408 (2025). https://doi.org/10.1007/s12539-024-00675-2
18. Liu, Y., Lian, Z.: PSDNet: A balanced architecture of accuracy and parameters for semantic segmentation. In: ICPR 2020, pp. 827–834 (2021). https://doi.org/10.1109/ICPR48806.2021.9412412
19. Pawar, R., Mukherji, P.: Mann-Whitney U test based identification of significant ECG features for improving accuracy of major depressive disorder classification. In: APCI 2025, pp. 1–6 (2025). https://doi.org/10.1109/APCI65531.2025.11136866
20. Davari, M.R., et al.: Probing representation forgetting in supervised and unsupervised continual learning. arXiv preprint arXiv:2203.13381 (2022)
21. Hassan, S., Rasheed, N., Qureshi, M.A.: A new regularization-based continual learning framework. In: HITE 2024, pp. 1–5 (2024). https://doi.org/10.1109/HITE63532.2024.10777234
22. Gunasekara, N., Pfahringer, B., Gomes, H.M., Bifet, A.: Survey on online streaming continual learning. In: IJCAI 2023, pp. 6628–6637 (2023). https://doi.org/10.24963/ijcai.2023/743
23. van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nat. Mach. Intell. **4**(12), 1185–1197 (2022). https://doi.org/10.1038/s42256-022-00568-3
24. Dalle Pezze, D., Anello, E., Masiero, C., Susto, G.A.: Continual learning approaches for anomaly detection. Evol. Syst. **16**(4), 111 (2025). https://doi.org/10.1007/s12530-025-09732-7

Analysis and Improvement of the Requirements Elicitation Process Based on ISO 31000

Fernanda Lopes de Lima[1(✉)], Simone Borges Simão Monteiro[1], and Viviane Vasconcellos Ferreira Grubisic[2]

[1] Departamento de Ciência da Computação, Universidade de Brasília (UnB), Brasília, Distrito Federal, Brasil
felima0730@gmail.com, simoneborges@unb.br

[2] Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina, Brasil
vivianegrubisic@unb.br

Abstract. Requirement elicitation is one of the most critical stages in the software development process, demanding a high level of accuracy as it aims to identify and thoroughly understand customer needs. This article seeks to analyze the current requirement elicitation process in a company operating in the information technology development sector. Based on this analysis, the study aims to identify, evaluate, and mitigate the risks involved, following the guidelines of the ISO 31000 standard, to support the organization in its operational and strategic decisions. To achieve the proposed objectives, an exploratory research methodology was adopted, applied in nature, and using a qualitative approach. Data collection techniques included document analysis within the company and brainstorming sessions. The results enabled the identification of the main risks associated with the requirement elicitation process. Finally, this research proposed a redesign of the elicitation model (TO-BE), resulting in significant improvements in the practices of the studied organization.

Keywords: Requirements Elicitation · Software Development · ISO 31000 · Risk Mitigation

1 Introduction

Software development is a complex activity, involving numerous unpredictable and difficult-to-control factors, such as technological innovations and constant changes in client requirements. This complexity causes a large portion of software development projects to exceed their planned schedule and budget, in addition to failing to meet client expectations in terms of functionality and quality [1].

Requirements elicitation techniques are essential to support requirements engineers in gaining a better understanding of the needs of users and stakeholders [14].

Given this scenario, in recent years there has been a growing interest among IT companies in the use of agile methodologies, with the primary objective of minimizing the problems constantly encountered in software development, such as: poorly defined

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 309–328, 2027.
https://doi.org/10.1007/978-3-032-30494-0_20

scope, requirement changes that directly impact the project's schedule and cost, and failure to meet client expectations.

It is also known that agile methodologies are innovative and iterative methods that support change and offer improvements in various aspects of systems and application development projects. However, their use in an unstructured and/or incomplete manner involves risks, which are greater for some organizations and projects [2].

According to research conducted by the Standish Group, approximately 43% of technology development projects tend to fail. Studies have shown that requirements defects are among the main sources of failure, accounting for 32.65%. This aspect has been neglected in requirements engineering and is generally considered a potential problem that can negatively affect projects [3].

A software requirement is a system property that must be identified and detailed by the analyst with the utmost precision and assertiveness. These requirements aim to automate parts of tasks, support an organization's business processes, correct flaws in existing software, or control devices. Interactions between users, business processes, and systems are often inherently complex. Therefore, risk management becomes essential in the requirements elicitation phase, considering that it involves a complex combination of contributions from various individuals working at different levels within an organization who are, in some way, directly involved in or connected to the environment in which the software will be implemented and used [2].

According to ISO 31000 – Risk Management, risk is the effect of uncertainty on objectives. The ISO also defines risk management as a set of coordinated activities to direct and control an organization with regard to risks and to create and protect value [4].

In this context, this research aims to apply the requirements elicitation process based on the ISO 31000 guidelines, with emphasis on risk identification, analysis, evaluation, and treatment. The goal is thereby to contribute to the company and those involved, offering a more assertive work process that adds value to the client, aligned with the use of agile methods.

This article is structured in six sections. The first presents a brief introduction that contextualizes the topic within the proposed research. The second section explores the most relevant works found in the literature; using the adapted TEMAC, sources related to the difficulties encountered in the requirements elicitation phase were identified. The third section provides a literature review, highlighting relevant authors on ISO risk management in the requirements elicitation process, as well as the Failure Mode and Effect Analysis (FMEA) tool. The fourth section addresses the methodology applied in the research, detailing the methods and structure adopted. The fifth section presents the results obtained. Finally, the sixth and last section proposes improvements to the requirements elicitation process and presents the final considerations.

To carry out a literature review, an adaptation of TEMAC was proposed with the objective of providing a list of important authors on the researched topic. For this purpose, the indexed databases Scopus and Web of Science were used. Initially, the search strings Requirements Elicitation, Risk management, Agile, and software development were applied to the filters. The period applied in the filter was the last ten years, in order to ensure that the observed facts were the most current within the literature. The

search on the Web of Science database returned 23 documents, including articles and journals, and 11 on the Scopus database. Exclusion was based on repeated articles and those classified in the research area of “Software design”, as well as terms outside the scope of the research, such as “Project management”, among others. As a result, 8 and 5 remained, respectively, totaling 13. After this bibliographic refinement process, this information was applied in the VOSviewer software, aiming to generate the visualization of bibliometric networks, where it is possible to verify the terms with the highest occurrence and relevance in the related articles.

Fig. 1. Bibliometric Network (Source: Elaborated by the authors, 2024)

The keywords found in the bibliometric network presented in Fig. 1 relate the main term "Requirements Engineering", followed by two equally important terms: Risk management and Requirements Elicitation, which are central subjects of this research. The term, whose translation is requirements engineering and requirements elicitation, was analyzed in the most widely used studies as research sources, that is, the main literature on the central topic of this study.

It is important to note that there are not many works related to risk management in the requirements elicitation phase available in the scientific databases searched. The largest collection found is linked to risk analysis in the software development process. However, the references cited above contain important information that contributed significantly to the development of this study.

2 Literature Review

2.1 Process Management

To support the requirements elicitation process, it is possible to use some process mapping methods for the area or even for the organization that requires an automated solution to improve company operations. One example is the Business Process Model and Notation (BPMN), which allows, through process notation, documenting the current workflow of the requirements elicitation process or even enabling redesign in order to propose improvements to the existing process. Many organizations use business process models to document business operations and to formalize these requirements in software engineering projects [5].

The use of BPM aims to understand and explore the processes carried out in an organization, measure them, and provide continuous improvement cycles for the organization's evolution. Along these lines, process mapping consists of the analysis ("AS-IS"), providing a detailed view of the process in question, showing the relationship between inputs, outputs, and the tasks performed. There are various ways to externalize and document the intellectual process for business process modeling, such as:

- Business Process Model and Notation (BPMN): A standard created by the Object Management Group (OMG®), approved as an American National Standards Institute (ANSI) standard, includes a simple and limited set of non-standardized symbols [6].
- Event-driven Process Chain (EPC): Considers events as triggers for a process step.
- Unified Modeling Language (UML): maintained by the OMG®, consists of a standard set of technical diagramming notations oriented toward describing the requirements of information systems.

There is a growing trend in the importance of modeling. Specification languages such as UML and BPMN are applied in the majority of software engineering projects today. However, each one has its application at specific stages of the software engineering process. When studying some works on the use of BPM in requirements elicitation processes, it was identified that one of the most widely used notations is BPMN, when the focus is on modeling the software process as a whole. The methodology for implementing process modeling in an organization is divided into four phases: Planning, Mapping and Analysis, Monitoring and Control, and Performance Evaluation. The proposal of this research is somewhat simpler, suggesting the identification, analysis, and evaluation of risks based on the current process (AS-IS) of requirements elicitation and the development of a process redesign proposal (TO-BE) for mitigating the identified risks [6].

2.2 SCRUM

In the game of Rugby, the "Scrum" is the way to restart the game after a foul. All players position themselves in a human huddle to compete for the ball. It is a play in which a team of eight members works together to move the ball forward in the opponent's field and achieve their single objective: the Goal.

The term Scrum was first associated with development by Hirotaka Takeuchi and Ikujiro Nonaka in the book "The New Product Development Game". The authors highlight the importance of adopting a development approach where the entire team works as a unit to achieve a common goal. In 1995, Ken Schwaber formalized Scrum for software development projects based on Toyota's Lean system [7].

In the Scrum framework, teams work as a highly integrated unit with each member playing a well-defined role and the entire team focusing on a single objective – product delivery. It eliminates unnecessary, inadequate, and bureaucratic control practices, concentrating on the essence of the process of building information systems.

Scrum is quite objective, having clear goals, a well-defined team, flexibility, commitment, and cooperation; and its learning curve is relatively low. Although Scrum is a methodology from the exact sciences field, it draws heavily from the humanities, and this must be taken into account at all times. Its pillars are:

Transparency. Significant aspects of the process must be visible to those responsible for the results. This transparency requires aspects defined by a common standard so that observers share a common understanding of what is being seen.

Inspection. Scrum users must frequently inspect Scrum artifacts and progress in order to detect variations. This inspection should not, however, be so frequent as to interfere with the actual execution of tasks. Inspections are most beneficial when performed diligently by inspectors specialized in the work being verified.

Adaptation. If an inspector determines that one or more aspects of a process have deviated outside acceptable limits, and that the resulting product would be unacceptable, the process or the material being produced must be adjusted. The adjustment must be made as quickly as possible to minimize further deviations.

According to its author Schwaber (2014), Scrum is not a predictable process; it does not define what to do in every circumstance. It is used in complex work where it is not possible to predict events and offers a framework and a set of practices that makes everything visible. This allows the team to have an accurate view of the facts throughout the project and, if necessary, make the appropriate adjustments to achieve their goals. This is one of Scrum's strengths: "Adaptability and Flexibility" [8].

Scrum is not a "miracle worker" and much less will it offer a "ready-made recipe" to solve all your problems. The only certainty is that these problems will be identified more easily. Being a framework, it will serve as a guide of best practices to achieve your objective.

Scrum prescribes four formal Events, contained within the limits of the Sprint, for inspection and adaptation:

- Sprint Planning Meeting – Objective: Held at the beginning of each Sprint to align everything that will be done so that the final product meets the needs of the end client.
- Daily Scrum Meeting – Objective: A short daily meeting for everyone to speak about the previous day, any problems that will need to be resolved, and what they intend to do that day.
- Sprint Review Meeting – The team involved in the project presents to the group the functionalities that were implemented during the focused Sprint. The Review is an accountability ceremony where the team presents the progress of this sprint to the Product Owner and other project stakeholders.
- Sprint Retrospective – At the restart of the cycle, using the learnings of this Sprint to plan the next one. It lasts on average one hour and takes place at the end of each Sprint.

3 Risk Management and ISO 31000

3.1 Risk Management

Risk Management is an essential component of project management, focused on identifying, evaluating, and responding to uncertainties that may affect project objectives. This edition adopts an approach based on principles and performance domains, emphasizing the importance of proactively managing uncertainties to optimize project outcomes [15].

The term risk is generally used to describe adverse events with a known probability, where uncertainty is found to be an element related to confidence in risk management and arises from a lack of knowledge and, therefore, has an unknown probability.

Risk can be considered anything that impacts an organization's objectives. Organizations of all types and sizes face internal and external influences and factors [2].

The 7th edition of the PMBOK® Guide highlights that risk management is not an isolated activity, but must be integrated throughout the entire project life cycle, promoting a culture of proactivity and continuous improvement. By adopting this approach, organizations can increase the probability of success of their projects, delivering consistent value and meeting stakeholder expectations [14].

One of the factors that can lead to the materialization of a risk is the exploitation of flaws or vulnerabilities, which often arise from the absence of adequate controls. Vulnerabilities consist of a set of weaknesses or internal factors that can be exploited by threats, generating the potential for incidents to occur, that is, undesired events [15].

Proactive risk management should be applied with the objective of protecting organizational assets—everything that has value to the organization, such as people, processes, environment, technology, information, and communication. This approach, in accordance with ISO 31000 guidelines, aims not only to mitigate risks, but also to strengthen organizational resilience and promote a strategic vision that supports the sustainability and success of the organization.

3.2 ISO 31000

ISO 31000 is an international standard developed by the International Organization for Standardization (ISO), which provides principles, guidelines, and a framework for risk management in any type of organization. The standard is widely recognized and used across various sectors as a reference for identifying, evaluating, treating, and monitoring risks in a systematic and effective manner [16].

The main characteristics of the ISO are:

- Objective: To help organizations manage risks proactively in order to achieve their strategic and operational objectives, create value, and protect assets and resources.
- Principle-Based Approach: The standard is built on principles that promote the integration of risk management into organizational culture and decision-making processes.
- Flexibility and Adaptability: Unlike certifiable standards, such as ISO 9001, ISO 31000 does not require strict compliance, but offers guidelines that can be adapted to the specific needs of each organization.
- Universal Applicability: It can be implemented in organizations of all sizes, sectors, and maturity levels in risk management.

ISO 31000 is structured in three main parts: Principles, Structure, and Process.

The first step of the process consists of establishing the context, which is divided into external and internal, listing a brief summary of the organizational objectives, since it is the risks of these objectives that, if not achieved, must be managed internally; the

organization has its organizational structure analyzed, as well as its responsibilities, processes, and systems. As for the external context, evaluations regarding the legal, social, financial, technological, and economic environments, among others, as well as the organization's relationship with stakeholders, also known as stakeholders, their perception and their values [2] (Fig. 2)

Fig. 2. Risk Management Process (Source: ISO 31000)

Identification of Risks. After establishing the context, it is necessary to identify **the** risks, that is, to find, recognize, and record the risks. This step must be based on the identification of risk sources, impact areas, events (including changes in circumstances), and their potential causes and consequences (Fig. 3).

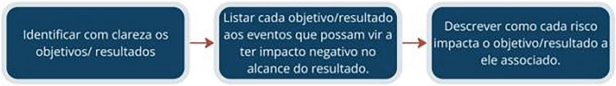


Fig. 3. Risk Identification (Source: Elaborated by the authors, 2024)

It is understood that the risk identification process is critical, given that an unmapped risk cannot be analyzed later, which can directly impact the failure of proactive risk management as a whole (Figs. 4 and 5).

Risk Analysis. Risk analysis refers to the development of an understanding of the risk and the determination of the risk level, as well as the measurement of the probability of the risk becoming an incident and the associated consequences.

Evaluation of Risks. Risk evaluation consists of comparing the estimated risk levels with risk criteria defined when the context was established, in order to determine the significance of the level and type of risks.

Fig. 4. Decisions in risk evaluation (Source: Adapted from ISO 31000)

Risk Treatment. Involves planning and carrying out actions to modify the risk level. The risk level can be modified through risk response measures that mitigate, transfer, or avoid these risks.

Fig. 5. Decisions in risk evaluation (Source: Elaborated by the authors, 2024)

Monitoring and Critical Review. Involves monitoring and verifying the performance or status of elements of risk management, which may encompass the policy, activities, risks, risk treatment plans, controls, and other matters of interest.

Continuous Improvement. Continuous improvement involves the enhancement or adjustment of aspects of risk management assessed during monitoring. It can be understood in two dimensions: one relating to the organization's own Risk Management System and another related to the monitoring results on the effectiveness of risk treatment, which is the responsibility of risk managers.

3.3 Failure Mode and Effect Analysis (FMEA)

Failure Mode and Effect Analysis (FMEA) was first developed as a formal method in design methodology in the 1960s by the aerospace industry due to its obvious reliability and safety requirements (SANKAR and PRABHU, 2001). Its use subsequently spread to other industries, such as automotive and oil and gas (PALADY, 1997). Failure Mode and Effect Analysis (FMEA) has long been used as a planning tool during the development of processes, products, and services. In the development of FMEA, the team identifies failures and actions that can reduce or eliminate the potential failure from occurring. Input is sought from a broad group of experts in design, testing, quality, product line, marketing, manufacturing, and the client to ensure that potential failure modes are identified.

FMEA is then used during the deployment of the product or service for troubleshooting and corrective action. The standard FMEA process evaluates failure modes for occurrence, severity, and detection. The multiplication of these values leads to what is known as the risk priority number [11].

The FMEA method translates, in a logical and systematic sequence, the evaluation of the possible ways in which a system or process is most subject to failure. FMEA considers the severity variables (S) of failures, the frequency with which they occur (O), and how they could be detected (D). Thus, based on these three indexes: severity, occurrence, and detection, the method leads to a prioritization of which failure modes of the product or service may cause the greatest risk to clients and to the organization itself [11].

From Table 1, it is possible to demonstrate how the severity, occurrence, and detection variables can be classified and related to each project, product, and industry.

Table 1. Severity and Failure Index (Source: Carbone, 2004)

Severidade e Índice de Falha		
Pontuação	**Descrição**	**Impacto no sistema ou processo**
10	Perigo sem aviso	1
9	Perigo com aviso	2
8	Funcionamento muito insatisfatório	3
7	Funcionamento insatisfatório	4
6	Redução moderada na capacidade do sistema	5
5	Impacto menor	6
4	Impacto mínimo	7
3	Leve incomodidade	8
2	Nenhuma consequência significativa	9
1	Nenhum impacto	10

Table 1 demonstrates the severity classification of a risk, where number 1 is considered very low and number 10 is highly dangerous according to the criterion. Table 2 demonstrates the classifications for the occurrence of failure or risk.

Table 2. Analysis Criteria for Failure Occurrence Index (Source: Adapted from FORD MOTOR COMPANY, 1988)

Critério de Análise para Índice de Ocorrência das Falhas		
Pontuação	**Descrição**	**Capacidade de detecção**
1	Falha sempre detectada	A falha é prontamente identificada
2	Alta probabilidade de detecção	A falha é facilmente detectável.
3	Boa probabilidade de detecção	A falha geralmente é identificada, mas nem sempre.
4	Média probabilidade de detecção	A falha pode ser detectada, mas depende de fatores.
5	Detecção ocasional	A falha é detectada em algumas situações.
6	Baixa probabilidade de detecção	A falha é raramente identificada.
7	Muito baixa probabilidade de detecção	A falha é detectada somente em casos excepcionais.
8	Dificuldade significativa para detectar	A falha é extremamente difícil de identificar.
9	Falha praticamente indetectável	É improvável que a falha seja detectada.
10	Falha impossível de detectar	A falha não pode ser detectada por nenhum meio.

In Table 2, it can be seen that the classification of the occurrence of a risk can be represented by a failure rate where 1 is considered almost impossible and 10 is extremely likely to occur. The classification criteria for Failure Detection are listed in Table 3, showing the failure indexes ranging from 1 to 10.

Thus, Table 3 highlights the classification regarding the possibility of detecting a failure or risk, ranging from 1, almost certain (easily detected), to 10, which are considered almost impossible to detect (difficult). The most widely used procedure to measure the risk associated with each failure mode is the multiplication of the scores obtained for the severity (S), occurrence (O), and detection (D) classifications. This results in a scale ranging from 1 to 1000 points, with 1 representing very low risk and 1000 a highly critical risk. This score is called the risk priority number (RPN), expressed as:

$$RPN = O \times S \times D$$

Table 3. Analysis Criteria for Failure Detection Index (Source: Adapted from FORD MOTOR COMPANY, 1988)

Critério de Análise para Índice de Ocorrência das Falhas		
Nível de Ocorrência	**Descrição da Taxa de Ocorrência**	**Índice de Ocorrência (O)**
Extremamente Alta	≥ 1 falha em 2 tentativas	10
Muito Alta	1 falha em 3 tentativas	9
Alta	1 falha em 8 tentativas	8
Moderada	1 falha em 20 tentativas	7
Significativa	1 falha em 80 tentativas	6
Baixa	1 falha em 400 tentativas	5
Muito baixa	1 falha em 2.000 tentativas	4
Rara	1 falha em 15.000 tentativas	3
Muito rara	1 falha em 150.000 tentativas	2
Quase impossível	1 falha em 1.500.000 tentativas	1

Where: RPN: Risk Priority Number, O: occurrence; S: severity; D: detection.

In this way, risk management, as a key component in the requirements elicitation process, through the use of the aforementioned methods and tools, will make it possible to establish a balance between the evaluated risks and their possible mitigation.

Based on the theoretical framework analyzed in this section, the methodology and research structure will be proposed in the next section.

4 Methodology

4.1 Research Method

The methodological proposal was based on conducting research with an exploratory objective, aiming to provide greater familiarity with the subject studied. The strategy employed was a case study, with an applied nature and a qualitative approach. According to Piovesan and Temporini (1995), this type of research aims to understand the variable being studied in its literal form, as it presents itself, along with its meaning and the context in which it is inserted. For this work, a Brainstorming model was followed, as it is a technique used to stimulate creative thinking and generate new ideas from a formal process that can be used in a structured or unstructured manner. Its main objective is to generate a high number of ideas in a creative and efficient way and, above all, free from criticism, and it is widely used in risk identification and problem solving (Fig. 6).

In order to propose a coherent study, we carried out the verification and analysis of the artifacts (documents) from the software development projects regarding the requirements elicitation phase, in addition to conducting Brainstorming sessions.

4.2 Research Structure

The research was structured in four stages, starting with the documentary research of the software development projects owned by the company. In the initial stage, potential risks were identified based on the current requirements elicitation process (AS-IS), along with the conduct of a Brainstorming session in which two professionals from the company

Fig. 6. Research Classification (Source: Elaborated by the authors, 2024)

who hold the position of project manager participated. The FMEA method was used, in which several process failures were presented [11].

After this study, the identification, analysis, evaluation, and treatment of the risks listed as critical in the process in question were carried out. Below is the figure showing the research structure of this study:

Fig. 7. Research Structure (Source: Elaborated by the authors, 2024)

At the end of this stage illustrated by Fig. 7, a process modeling technique was applied in Bizagi, based on the BPMN notation. In this way, it was possible to apply the risk management process, through the proposed research methodology and structure, which will be further detailed in Sect. 5.

5 Results and Discussion

5.1 Context Establishment

The company studied is a corporation, 100% Brazilian, and privately held. Founded in 2010, it operates in the IT market with a focus on developing customized solutions.

Its mission is to be a prosperous business group in the IT segment, providing success to clients, employees, and shareholders, through an innovative, genuine, and achievable vision. Its vision is to provide clients, through qualified human resources and powerful management tools, with the means to achieve their objectives while adding value to the business.

Its value chain is divided into: Strategic Processes, Core Processes, and Support Processes, as shown in Fig. 8

As the central object of this process illustrated in Fig. 8, with respect to the company's Value Chain, the end-purpose process has the macro-process "Develop Software". It

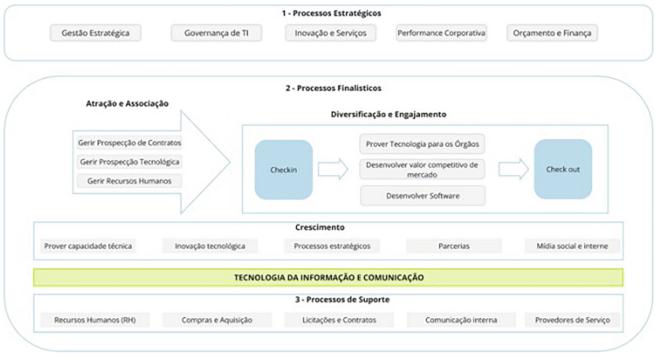


Fig. 8. Value Chain (Source: Elaborated by the authors, 2024)

was possible to detail and reach the requirements elicitation process, where a process is being used by the company, and it was possible to describe the main stages through the modeling of this process.

According to Fig. 9, it was possible to interpret the figure and identify the risks from the modeled process (AS-IS).

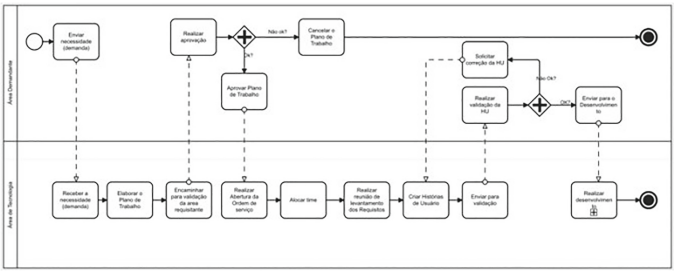


Fig. 9. AS-IS Mapping (Source: Elaborated by the authors, 2021)

5.2 Risk Management – Identification

A documentary research of the company's projects regarding the requirements elicitation phase represented by the process in Fig. 9 was conducted. Based on this study and the historical analysis of the projects, Brainstorming was carried out with the participation of two senior project managers in software project management. In this Brainstorming session, problems were identified in some of the events of the requirements elicitation process. After identifying these failures, it was possible to carry out the risk identification.

As shown in Table 4, the events that had one or more problems/failures and risks that were identified are listed.

Table 4. Risk Identification (Source: Prepared by the authors)

PROCESSO DE ELICITAÇÃO DE REQUISITOS		
Evento	**Problema**	**Risco Identificado**
Enviar necessidade (demanda)	• Atualmente não há uma ferramenta de gestão de controle de solicitações (demanda). Abertura de demanda ocorre atualmente via e-mail e Teams.	• Falta de padronização de atendimento, ausência de controle das solicitações, além da descentralização das informações.
Receber a necessidade (demanda)	• Por ter mais de um canal de entrada de solicitação, pessoas diferentes atuam na mesma demanda. Causando duplicidade e retrabalho.	• Retrabalho, perda de prazos junto aos clientes, causando insatisfação é um sentimento de desorganização e amadorismo por parte da empresa.
Elaborar Plano de Trabalho	• Não há muito detalhamento para elaboração do documento que estima prazo e custo.	• Por vezes não há detalhamento na solicitação para que seja criada uma proposta com escopo definido, orçamento e prazo factível de ser executado. Além disso não há padronização do formato do documento e nem em sua escrita. Esses problemas causam dúvidas e falta de clareza no plano de trabalho elaborado pela empresa.
Encaminhar para validação da área requisitante	• Caso a área requisitante não concorde com o Plano de trabalho poderá realizar o cancelamento sem nenhum ônus. Com isso a empresa não é remunerada pelo trabalho realizado.	• Todo o trabalho realizado e as horas gastas pelo ou pelos profissionais que elaboraram o Plano de trabalho não é pago pela contratante. Causando prejuízo à empresa.
Alocar Time	• A empresa não possui squads ou equipes dedicadas para lidar com diferentes tipos de projetos, operando exclusivamente sob demanda. Sempre que uma ordem de serviço é aberta, os integrantes do time são selecionados sem critérios claros. Essa abordagem apresenta sérios problemas, como dificuldades na curva de aprendizado, ausência de gestão estratégica de competências (skills) dos profissionais e aumento do turnover, causado pela desmotivação decorrente da falta de estrutura e alinhamento.	• Baixa Eficiência Operacional • Dificuldade na Curva de Aprendizado • Falta de Continuidade e Conhecimento Acumulado • Falta de Gestão de Competências (Skills) • Desmotivação e Turnover Elevado • Impactos na Qualidade das Entregas • Aumento de Conflitos Internos • Riscos Financeiros e Atrasos • Dificuldade na Gestão de Recursos • Falta de Inovação e Crescimento
Realizar reunião de levantamento dos Requisitos	• Ausência de processo e método na etapa de descoberta. • Excesso de atividades para os analistas de requisitos • Ausência de repasse de conhecimento dos projetos aos analistas • Falta de um plano de comunicação apontando os envolvidos e os responsáveis pelo projeto • Ausência de uma ferramenta de acompanhamento e repositorio dos artefatos elaborados • Ausência de uma matriz de rastreabilidade dos requisitos mais complexos.	• Levantamento inadequado ou incompleto de requisitos, resultando em escopo mal definido. • Falhas na identificação de stakeholders e suas necessidades. • Aumento de retrabalho devido à descoberta tardia de informações críticas. • Dificuldade para alinhar expectativas e prioridades entre a equipe e os clientes. • Sobrecarga de trabalho, levando à queda na qualidade das análises e entregas. • Atrasos no cronograma devido à baixa capacidade de entrega. • Desmotivação e aumento do turnover dos analistas devido ao estresse. • Erros e omissões nos requisitos, impactando diretamente no desenvolvimento. • Aumento do tempo necessário para que novos analistas atinjam produtividade. • Dependência de poucos profissionais, criando vulnerabilidade em caso de turnover. • Diminuição da colaboração e alinhamento entre os times. • Aumento do risco de não conformidade com as expectativas dos stakeholders ou reguladores. • Dificuldade em identificar impactos de mudanças nos requisitos, resultando em inconsistências no sistema. • Falhas na priorização dos requisitos, comprometendo a entrega de valor.

From the identification and analysis of the workflow scenarios, it was possible to identify the problems, and from this the related risks were reached.

5.3 Risk Management – Analysis and Evaluation

In order to propose the analysis and evaluation of risks, FMEA was used, through the analysis of failure modes, which translates into reliability and quality control, as the documentation enables the organization to recognize potential failures in the process in advance, and if they happen to occur, have sufficient tools to resolve them in the best way possible. For the construction of the failure modes, the main risks described in Table 5 were considered:

Table 5. Failure Mode and Effect Analysis (Source: Prepared by the authors)

Análise de Riscos - FMEA								
Etapa do Processo	**Risco Identificado**	**Efeito Potencial do risco**	**Grau de Severidade (1-10)**	**Índice de Ocorrência (1-10)**	**Índice de Detecção (1-10)**	**RPN (Resultado)**	**Ações Recomendadas**	**Responsável**	**Prazo**
Enviar necessidade (demanda)	Falta de padronização de atendimento, ausência de controle das solicitações, além da descentralização das informações.	Solicitações tratadas de forma inconsistente, resultando em atrasos, erros e falta de rastreabilidade e falta de centralização das informações.	6	7	7	441	Implementar uma ferramenta de gestão de demandas centralizada para registrar, monitorar e rastrear solicitações. Definir processos padronizados de atendimento para garantir consistência e clareza nas entregas.	Coordenação da Tecnologia de Informação	45 dias
Alocar Time	Desmotivação e Turnover elevado	Prejuízos financeiro e de conhecimento de negócios para a empresa	7	6	7	294	Desenvolver uma estratégia de retenção de conhecimento que inclua documentação robusta e repositórios acessíveis. Implementar mecanismos de governança para reduzir o impacto financeiro, como controle de custos e análise de retorno de investimento.	Coordenação da Tecnologia de Informação e RH	90 dias
Alocar Time	Impactos na Qualidade das Entregas	Produtos ou serviços entregues com defeitos, não atendendo aos requisitos dos clientes ou stakeholders, resultando em retrabalho, insatisfação do cliente e possíveis penalidades contratuais.	7	8	6	336	Realizar testes de validação e verificação durante todas as etapas do projeto para garantir que os produtos atendam aos requisitos. Criar um plano de garantia de qualidade (QA), incluindo critérios claros de aceitação.	Equipe técnica dos projetos	30 dias
Realizar reunião de levantamento de requisitos	Levantamento inadequado ou incompleto de requisitos, resultando em escopo mal definido	Escopo imprevisto leva a entregas desalinhadas com as necessidades reais do projeto, retrabalho constante e aumento do custo e do tempo de execução.	9	8	8	648	Adotar **métodos ágeis** para garantir um ciclo contínuo de feedback e adaptação do escopo. Realizar **workshops de levantamento de requisitos** com a participação dos stakeholders para alinhamento claro.	Coordenação da Tecnologia de Informação	30 dias
Realizar reunião de levantamento de requisitos	Ausência de metodologia ágil	Lentidão no processo de adaptação a mudanças, dificuldades em priorizar entregas de valor e maior rigidez na gestão do projeto, comprometendo a eficiência e a satisfação dos stakeholders.	9	9	7	567	Introduzir práticas de **gestão ágil**, como Scrum ou Kanban, para permitir maior flexibilidade. Implementar **priorização baseada em valor**, utilizando técnicas como MoSCoW ou análise de valor agregado. Estabelecer uma **cultura organizacional aberta a mudanças**, com treinamento contínuo das equipes.	Diretoria da Tecnologia da Informação	30 dias
Realizar reunião de levantamento de requisitos	Aumento de retrabalho devido à descoberta tardia de informações críticas.	Retrabalho elevado aumenta os custos, causa atrasos e pode gerar insatisfação tanto da equipe quanto dos stakeholders, além de comprometer o cumprimento de prazos.	8	6	6	288	Garantir que todos os requisitos sejam revisados e validados antes de iniciar o desenvolvimento. Criar checklists e templates padronizados para evitar inconsistências. Adotar práticas de gestão de riscos para identificar potenciais falhas precocemente.	Equipe técnica dos projetos	30 dias
Realizar reunião de levantamento de requisitos	Erros e omissões nos requisitos, impactando diretamente no desenvolvimento.	Desenvolvimentos incorretos ou incompletos, resultando em produtos defeituosos, custos adicionais com correções e atrasos na entrega final	6	5	6	180	Reforçar o Levantamento e Validação de Requisitos: Implementar Revisões e Testes ao Longo do Desenvolvimento. Adotar Metodologias Ágeis Melhorar a Comunicação e Documentação Capacitação e Treinamento da Equipe	Equipe técnica dos projetos	30 dias
Realizar reunião de levantamento de requisitos	Dificuldade em identificar impactos de mudanças nos requisitos, resultando em inconsistências no sistema.	Alterações mal gerenciadas criam falhas sistêmicas, perda de funcionalidade ou integração inadequada, comprometendo o desempenho do sistema como um todo.	5	4	6	120	Implementar Gerenciamento de Mudanças: Estabelecer a Rastreabilidade dos Requisitos: Analisar Impactos Antes de Aprovar Mudanças: Utilizar Ferramentas de Gerenciamento de Requisitos: Aplicar Métodos Ágeis e Incrementais:	Coordenação de TI Equipe técnica dos projetos	30 dias
Realizar reunião de levantamento de requisitos	Falhas na priorização dos requisitos, comprometendo a entrega de valor.	Requisitos de menor importância podem ser priorizados em detrimento de itens essenciais, resultando em entregas que não agregam valor significativo ao cliente ou ao projeto.	6	8	7	336	Aplicar Métodos de Priorização Estruturados Utilizar metodologias reconhecidas para priorizar os requisitos com base em critérios objetivos e alinhados ao valor de negócio. Exemplos incluem: - Matriz de Eisenhower: Prioriza requisitos com base em urgência e importância. - Método MoSCoW: Classificação dos requisitos.	Equipe técnica dos projetos	30 dias

For a better understanding of Table 5, a description of each field follows:

- Process Stage: Which stage of the project is being evaluated.
- Risk: Identification of the current risk.
- Potential Effect of the Problem: Expected impact if the problem occurs.
- Severity Level: Classification of the severity of the problem (1–10).
- Occurrence Index: Probability of the problem occurring (1–10).
- Detection Index: Probability of the problem being detected before causing impact (1–10).
- RPN (Result): Multiplication of the three indexes (Severity × Occurrence × Detection).
- Recommended Actions: Suggested measures to mitigate the identified risk.
- Responsible: Person or team responsible for implementing the actions.
- Deadline: Estimated time to carry out the proposed actions.

Based on these reports and the analysis of failure modes and effects, the Risk Priority Number (RPN) values were calculated, with the highest values highlighted in yellow rows. The RPN of 648 was assigned to the risk "Inadequate or incomplete requirements gathering, resulting in a poorly defined scope"; the risk "Absence of agile methodology" obtained an RPN of 567; and the risk "Lack of standardization in request handling, in addition to the decentralization of information" had an RPN of 441.

The objective of this stage was to define the level of the identified risks, as well as to prioritize them for treatment. As input in this process, risk records and the probability and impact matrix were used, which contains the score calculation for each risk. To define the risk level, three evaluation levels were used, namely:

- Range 01 (green color): represents the risks with the lowest level of relevance in relation to the analysis performed. For this level, the risk must have a score between 0.1 and 0.9.
- Range 02 (yellow color): represents the risks with an intermediate level of relevance in relation to the analysis performed, but which require a degree of attention for evaluation. For this level, the risk must have a score between 1 and 2.1.
- Range 03 (red color): represents the risks with a high level of relevance in relation to the analysis performed and indicates a critical point that must be considered in the risk evaluation. For this level, the risk must have a score between 2.5 and 4.5 [12].

For a better understanding, the method used will be represented in Table 6.

Table 6. Risk Level Table (Source: Adapted from ISO 31010)

PROBABILIDADE						
	90%	Média	Média	Média	R05 - Ausência de Métodos Ágeis	R04 - Levantamento de requisitos inadequado.
	70%	Baixa	Média	R3 - Impactos na Qualidade das Entregas	R9 - Falhas na priorização dos requisitos	R01 - Falta de padronização de atendimento (ausência de ferramenta).
	50%	Baixa	Baixa	Baixa	R7 - Erros e omissões nos requisitos, impactando diretamente no desenvolvimento.	R6 - Aumento de retrabalho devido à descoberta tardia de informações críticas.
	30%	Baixa	Baixa	Baixa	R8 - Dificuldade em identificar impactos de mudanças.	R9 - Falhas na priorização dos requisitos.
	10%	Baixa	Baixa	Baixa	Baixa	Média
		Muito Baixo	Baixo	Moderado	Alto	Muito Alto
		IMPACTO				

As presented in Table 6, three risks represent a high level of relevance in relation to the analysis performed and were indicated as critical in their evaluation; in yellow, six risks with an intermediate level of relevance in relation to the analysis performed were analyzed and evaluated.

Thus, Table 7 shows the evaluation and treatment plan for the identified risks.

Table 7. Risk Analysis and Evaluation (Source: Prepared by the authors)

Análise e Avaliação dos Riscos						
Risco	**Tipo**	**Probabilidade**	**Impacto**	**Ação**	**Medidas**	**Medida(s) Aprovada(s)**
Falta de padronização de atendimento, ausência de controle das solicitações, além da descentralização das informações.	Negativo	Alto	Alto	Implementar uma ferramenta de gestão de demandas centralizada para registrar, monitorar e rastrear solicitações. Definir processos padronizados de atendimento para garantir consistência e clareza nas entregas.	Elaborar proposta de redesenho do processo (TO-BE) Aquisição de uma ferramenta de gestão de demandas.	x
Desmotivação e Turnover elevado	Negativo	Alto	Moderado	Desenvolver uma estratégia de retenção de conhecimento que inclua documentação robusta e repositórios acessíveis. Implementar mecanismos de governança para reduzir o impacto financeiro, como controle de custos e análise de retorno de investimento.	Desenvolver uma estratégia de retenção de conhecimento com documentação eficiente e implementar mecanismos de governança para controle de custos e análise de retorno sobre investimento	x
Impactos na Qualidade das Entregas	Negativo	Alto	Moderado	Realizar testes de validação e verificação durante todas as etapas do projeto para garantir que os produtos atendam aos requisitos. Criar um plano de garantia de qualidade (QA), incluindo critérios claros de aceitação.	Realizar testes de validação e verificação em todas as etapas do projeto e criar um plano de garantia de qualidade (QA) com critérios claros de aceitação.	
Levantamento inadequado ou incompleto de requisitos, resultando em escopo mal definido.	Negativo	Alto	Alto	Adotar **métodos ágeis** para garantir um ciclo contínuo de feedback e adaptação do escopo. Realizar **workshops de levantamento de requisitos** com a participação dos stakeholders para alinhamento claro.	Elaborar uma proposta de redesenho do processo (TO-BE) Realizar a adoção de algum framework ágil. Realizar a implantação do Upstream nos projetos.	x
Ausência de metodologia ágil	Negativo	Alto	Alto	Produzir práticas de **gestão ágil**, como Scrum ou Kanban, para permitir maior flexibilidade. Implementar **priorização baseada em valor**, utilizando técnicas como MoSCoW ou análise de valor agregado. Estabelecer uma **cultura organizacional aberta a mudanças**, com treinamento contínuo das equipes.	Elaborar uma proposta de redesenho do processo (TO-BE) Realizar uma transformação ágil na empresa	x
Aumento de retrabalho devido à descoberta tardia de informações críticas.	Negativo	Alto	Moderado	Garantir que todos os requisitos sejam revisados e validados antes de iniciar o desenvolvimento. Criar checklists e templates padronizados para evitar inconsistências. Adotar práticas de gestão de riscos para identificar potenciais falhas precocemente.	Implantação de Upstream na etapa de levantamento de requisitos. Aquisição de ferramenta de controle e execução dos projetos.	
Erros e omissões nos requisitos, impactando diretamente no desenvolvimento.	Negativo	Alto	Alto	Reforçar o Levantamento e Validação de Requisitos; Implementar Revisões e Testes ao Longo do Desenvolvimento; Adotar Metodologias Ágeis; Melhorar a Comunicação e Documentação Capacitação e Treinamento da Equipe	Implantação de Upstream na etapa de levantamento de requisitos. Aquisição de ferramenta de controle e execução dos projetos.	
Dificuldade em identificar impactos de mudanças nos requisitos, resultando em inconsistências no sistema.	Negativo	Alto	Moderado	Implementar Gerenciamento de Mudanças; Estabelecer a Rastreabilidade dos Requisitos; Analisar Impactos Antes de Aprovar Mudanças; Utilizar Ferramentas de Gerenciamento de Requisitos; Aplicar Métodos Ágeis e Incrementais;	Aquisição de uma ferramenta (Wiki) onde possa ser centralizada as informações, documentos dos projetos.	x
Falhas na priorização dos requisitos, comprometendo a entrega de valor.	Negativo	Alto	Moderado	Aplicar Métodos de Priorização Estruturados Utilizar metodologias reconhecidas para priorizar os requisitos com base em critérios objetivos e alinhados ao valor de negócio. Exemplos incluem: • Matriz de Eisenhower: Prioriza requisitos com base em urgência e importância. • Método MoSCoW: Classificação dos requisitos.	Realização de transformação ágil. Capacitação dos profissionais Plano de ação junto ao RH para Política de Certificações aos profissionais.	x

Based on this evaluation, it was possible to assist the company studied, identifying the highest risks that should have a decision made in response to them.

5.4 Improvement Proposal

As a way to mitigate or even eliminate the risks identified in the requirements elicitation process (AS-IS), the modeling of the process redesign in question (TO-BE) was proposed, in order to optimize the previous process and resolve various failures identified in the process.

The improvement of the current requirements elicitation process is based on good risk management and the quality and continuous improvement process, minimizing the risk factors such as:

- F01 – Lack of effective communication among stakeholders (clients, analysts, developers).
- F02 – Inexperience or lack of training in the requirements analysis team.
- F03 – Lack of structured agile tools and methods to capture and document requirements.
- F04 – Limited or nonexistent participation of stakeholders in the elicitation process.
- F05 – Insufficient deadlines or pressure to accelerate this critical stage of the project.

These threats can cause negative impacts on organizational objectives, as listed below:

- R04 – Inadequate requirements gathering.
- R05 – Absence of Agile Methodologies.
- R01 – Lack of standardization in service delivery (Absence of a tool).

Based on the risks, their factors and impact, focusing on the quality of requirements elicitation, prioritizing the mitigation and elimination of risks, and continuous improvement in the process, the TO-BE process represented by Fig. 10 was created.

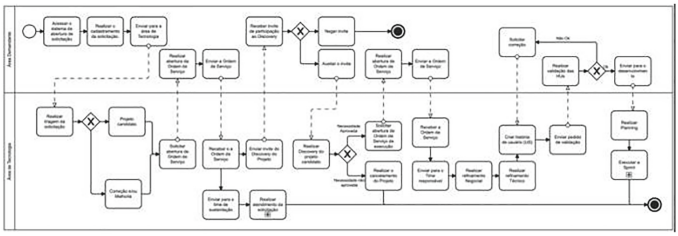


Fig. 10. TO-BE Mapping (Source: Elaborated by the authors, 2021)

Through the TO-BE approach, an optimized strategy was proposed for the acquisition and use of a centralized tool for opening service requests, which eliminated the duplication of entries coming from various channels in the company's technology area. This change brought greater control and traceability, reduced operational failures, and optimized the workflow.

In the proposed solution, the company adopted a demand management tool that, upon receiving requests from end-user requesters, automatically performed a triage based on the type of request. This functionality allowed for the creation of an area dedicated exclusively to supporting legacy systems, which optimized the service delivery process. As a result, professionals were able to work with greater focus and peace of mind in developing their activities, which increased productivity and the quality of the delivered results.

Another fundamental point was the implementation of an agile methodology, with a focus on the requirements gathering phase, introducing the Discovery practice in the requirements elicitation phase (Upstream). This approach enabled a deeper understanding of the needs on the part of the intervening areas and the technology team, promoting alignment from the very beginning of the project.

Important new stages were incorporated into the process, such as:

Business Refinement: In this phase, the Product Owner (PO) carries out a detailed survey of all business rules together with the requesting area, ensuring clarity and completeness.

Technical Refinement: Carried out in collaboration with the technical part of the team, that is, the Dev Team. This step allows for the validation of technical aspects and the feasibility of requests, aligning technical expectations with business needs.

After the completion of the refinements, it was possible to create the User Stories, ensuring that the requirements were properly defined. This practice played a crucial

role in mitigating risks related to incomplete or incorrect requirements, contributing significantly to improving the quality of deliverables.

In the TO-BE process, one of the main proposed activities was the organization of the Service Order (SO) opening flow, which previously lacked a formal process. The absence of this structure caused losses for the company, since some work performed was not formalized, which resulted in a lack of payment from the client area.

Regarding the processes that precede requirements elicitation, it was defined that they will be reviewed and redesigned to address the identified gaps. One example is the improvement of the process for signing and recording Service Orders. It was established that the prerequisite for executing any activity linked to a project will be the formal opening and signing of the Service Order, ensuring alignment between the parties and the formalization of commitments.

The agile methodology adopted by the teams was SCRUM, with all ceremonies being followed rigorously. This implementation brought greater efficiency, providing more assertive deliveries, both in terms of the quality expected by the client and in meeting the established deadlines. After the elicitation and approval of the requirement by the requesting area, it was forwarded to the Product Owner (PO) responsible for the development team. The PO, together with the team, performed the Sprint Planning ceremony, ensuring the execution of the stories included in the sprint.

To mitigate the risks of requirement changes during or after development, a Squad of test analysts (QA) was created to assist the PO in testing. This team was responsible for validating the User Stories (US), ensuring that deliveries were aligned with the specified requirements, reducing errors and rework.

Another relevant point was the adoption of a tool that streamlined the work of the development team: Trello. This tool provided greater control over the execution of the user stories included in the sprint, in addition to facilitating the management of the demands of the support team.

Additionally, team integration, combined with the use of appropriate tools and methodologies, raised the level of organization, predictability, and stakeholder satisfaction in the software development process. These improvements promoted a collaborative and efficient environment, ensuring value delivery to the client and the success of the projects.

6 Conclusion

Our objective was achieved, as we sought to analyze the risks in the requirements elicitation phase and propose improvements to the current process of the company in question.

We concluded, through this study, that it is possible to manage risks during the requirements elicitation process. We performed a detailed analysis of the current process (AS-IS) to identify, analyze, evaluate, and treat the risks involved. During this analysis, we identified critical points that highlighted the significant need for efficient risk management.

To this end, we developed and applied a risk management model specific to the elicitation phase. We used various mechanisms, such as benchmarking with company

experts and the analysis of documented projects, to identify the risks involved. Based on these efforts, we were able to propose contingency actions that could be implemented as internal controls.

In addition, we conducted a bibliographic review to explore possible methods of risk analysis and evaluation that could be applied in the context of requirements elicitation. In response to the risks identified in the AS-IS process, we recognized the need to redesign the process to an improved state (TO-BE), which was redesigned and presented to the company. It is important to highlight that the organization's needs are dynamic and evolving.

Therefore, we believe that the process redesign can be constantly monitored, evaluated, and improved, ensuring efficient risk management in the requirements elicitation phase.

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

References

1. Pressman, R.M.: Software Engineering, 8th edn. McGraw Hill Brasil (2016)
2. Chandani, P., Gupta, C.: An early multi-criteria risk assessment model: requirement engineering perspective. J. Cases Inform. Technol. **21**(2), 51–64 (2019)
3. Standish Group: Chaos Report. https://www.standishgroup.com/
4. Brazilian Association of Technical Standards: ABNT ISO GUIDE: Risk Management – Principles and Guidelines. ABNT, Rio de Janeiro (2009)
5. ABNT: NBR ISO IEC 31010:2009 – Risk management — Risk management – Techniques for the risk assessment process. ABNT, São Paulo (2012). ABPM International: Guide to the Business Process Management Body of Knowledge (BPM CBOK®). ABPMP (2020)
6. Object Management Group OMG®: Business Process Model and Notation (BPMN) (2010). http://www.omg.org/spec/BPMN/2.0
7. Sutherland, J.: SCRUM: The Art of Doing Twice the Work in Half the Time. Leya (2014)
8. Machado, M., Medina, S.G.: SCRUM–Agile Method: a cultural change in Software Development Project Management. Revista Científica Intraciência. Faculdade do Guarujá–UNIEESP **1**(1), 58–71 (2009)
9. Moraes, G.M.D.: Process analysis and improvement in software reengineering: a proposal for a process risk management instrument (2021)
10. Ministry of Transparency and Federal Comptroller General: Risk management methodologies (2018). https://www.gov.br/cgu/pt-br/centrais-deconteudo/publicacoes/publicacoes-institucionais/arquivos/cgu-metodologia-gestaoriscos2018.pdf
11. Carbone, T.A., Tippett, D.D.: Project risk management using the project risk FMEA. Eng. Manag. J. **16**(4), 28–35 (2004)
12. M.I.C. et al.: How to elicit and specify software requirements from bpmn diagrams? In: Proceedings of the XIV Brazilian Symposium on Information Systems, pp. 1–8 (2018)
13. Canedo, E.: Guide to requirements elicitation techniques applied to agile software development
14. Garcia, V.M.B.: The PMBOK guide and the changes in the 7th Edition. Revista Inovação, Projetos e Tecnologias **10**(1), 123–125 (2022)

15. de Almeida Araújo, M.F., et al.: Requirements engineering risks with the use of agile methodologies
16. Vieira, A.A.C., do Amaral, N.: RISK MANAGEMENT–ISO 31000. Associação de Ensino Julian Carvalho–AEJC, 54

General Track 6: Urban and Regional Planning

A Structured Comparative Review of Digital Structural Health Monitoring Solutions

Lambros Mitropoulos and Eirini Stavropoulou(✉)

School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Zografou Campus, 15780 Athens, Greece
{lmit,eirini_stavropoulou}@mail.ntua.gr

Abstract. The rapid digitalization of transport infrastructure has led to the development of heterogeneous tools for Structural Health Monitoring (SHM), damage detection, and predictive maintenance. However, the lack of harmonized descriptors and comparative frameworks limits systematic comparison, transferability, and evidence-based decision making. This paper introduces a novel comparative framework for the structured synthesis and cross-solution analysis of existing SHM tools, based on a harmonized metadata model and similarity-based analysis. This framework is applied to thirteen representative tools, including spanning vision-based inspection, sensor-driven SHM, subsurface assessment, and commercial inspection systems. The analysis identifies dominant solution groups, maturity patterns, and recurring deployment barriers. Results show a strong concentration on monitoring and detection capabilities, limited integration of control or closed-loop interventions, and persistent gaps related to data availability, multimodal integration, and standardization. The proposed approach demonstrates how structured metadata and similarity analysis can support knowledge harmonization and inform future research and deployment strategies in infrastructure health management.

Keywords: Structural Health Monitoring (SHM) · Infrastructure Health Monitoring (IHM) · health monitoring · digital inspection · predictive maintenance · smart infrastructure

1 Introduction

Transport infrastructure stability and resilience are necessary to maintain road safety, service continuity, and sustainable functionality of mobility systems [1]. Assessing road sections accurately is important in preventing hazardous conditions, mitigating risks, and allocating resources effectively [2]. Advances in sensory technology, data processing, and digital platforms have introduced multiple ways for implementing Structural Health Monitoring (SHM) [3], which has a significant role in ensuring sustainability and safety of global civil infrastructure networks [4]. These include various techniques such as vision-based damage detection, vibration-based SHM, digital twin enabled assessment, and predictive maintenance applications [3, 5–10]. Smart city initiatives and intelligent transportation systems increasingly integrate these technologies into critical assets like

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 331–345, 2027.
https://doi.org/10.1007/978-3-032-30494-0_21

roads and bridges [11]. Additionally, artificial intelligence (AI) enables the use of continuous SHM to proactively identify issues before they become major problems, enabling predictive maintenance [12–14].

Although rapid advancements in technological capabilities exist, current landscape of digital SHM solutions remains highly fragmented. Each solution utilizes different sensor modalities, analytical approaches, spatial scope, maturity levels, and validation contexts [15], making meaningful comparisons difficult for researchers seeking to synthesize state of the art and practitioners searching for suitable technologies to deploy. Additional complications also exist due to the lack of consistent terminology and structured, comparative information regarding solution characteristics, evidence strength, and deployment barriers [11, 16, 17]. Energy availability and data limitations continue to hinder scale-up of Internet of Things (IoT) technologies used for continuous sensing and monitoring [11].

This study provides a structured synthesis and comparative analysis of existing digital solutions rather than assessing new monitoring technologies. The objective is to organize a representative set of SHM tools within a common framework, utilizing harmonized descriptors to highlight dominant technical patterns, functional groupings, and recurring challenges within the field [9]. This initiative aligns with European research and innovation efforts, such as the IM-SAFE action and other projects like Sustainable Bridges, LIFECON and SMARTIN which support knowledge harmonization, capacity building, and evidence-informed infrastructure management [18]. Through this structured and practitioner oriented perspective, the paper aims to clarify how current SHM solutions relate to one another, where robust evidence exists, and which gaps continue to constrain broader implementation.

2 Literature Review

SHM involves continuous or periodic measurements of structural behavior, environmental conditions, and operational loads to determine its condition and predict its remaining service life [19]. Four levels are common in the SHM process. These are: Damage detection (identifies the existence of damage); Damage localization (locates where the damage exists); Damage classification (determines what kind of damage has occurred and how severe it is); and Damage prognosis (predicts when the structure will no longer be able to perform its intended function) [20, 21].

Most traditional methods used in SHM involve visual inspection, analytical models, and signal processing techniques [22]. While these have been successful in identifying some forms of damage, they have several limitations in detecting internal damage, predicting complex structural behaviors under various loading conditions, and analyzing large amounts of data generated by distributed sensor networks [23]. Recent advances in data-driven, machine learning approaches has addressed many of these limitations, enabling automated feature extraction, pattern recognition, and predictive analytics [24, 25].

Deep learning is rapidly becoming the most popular method of analyzing data generated by SHM systems. Due to the ability to recognize patterns in large datasets, deep learning has outperformed traditional machine learning approaches related to damage

detection, localization, and classification [20, 26]. Cha et al. [20] provide a comprehensive review of deep learning-based SHM, encompassing non-destructive approaches, computer vision-based methods, digital twins, and vibration-based strategies. They also establish connections between traditional machine learning and deep learning methods, highlighting the evolution toward end-to-end learning systems [20].

One major advantage of unsupervised learning methods is that they do not require the availability of labeled damage data from real structures, which is a major problem in SHM [27]. As a result, vibration-based SHM has benefited greatly from unsupervised approaches. These approaches create baseline models of normal structural behavior and then use them to detect any deviation in that behavior that could indicate damage [27].

Lawal et al. [28] integrated AI into wireless smart sensor platforms for railroad bridge impact detection, implementing machine learning predictions directly on Xnode wireless sensor nodes [28]. This allows for rapid assessment and response during time-sensitive applications such as detecting impacts from over-height vehicles.

As AI becomes more prevalent in SHM, questions about model interpretability, trustworthiness, and credibility of predictions arise [29]. Mondal & Chen [29] discuss contemporary issues regarding explainable physics-informed AI in civil SHM, while outlining future research directions for innovative AI applications. Developing interpretable machine learning models that incorporate principles of structural mechanics will enhance stakeholder confidence and regulatory acceptance of AI-based SHM systems [30].

Another critical challenge is the effective integration of SHM data with existing asset management systems, maintenance planning processes, and decision-making frameworks [31]. The gap between research innovations and practical implementation in transportation agencies is explained by organizational, regulatory, and technical barriers [32]. Therefore, standardized data formats, interoperable systems, and decision support tools must be developed to realize the full potential of SHM technologies [31].

Long-term sensor reliability, data corruption, and communication failures are other significant challenges for SHM systems [33]. Wireless Sensor Networks (WSNs) limitations, particularly energy supply constraints, affect system sustainability and maintenance requirements [33]. Development of energy-autonomous sensors with energy harvesting capabilities addresses some concerns, however ensuring reliable operation over infrastructure design life times (50–100 years) remains challenging [34].

The perspective of this paper aligns closely with the ambitions of the SMARTIN project, which seeks to accelerate the digital transformation of European transport infrastructure by consolidating existing tools, promoting interoperability, and enabling data-driven asset management. SMARTIN emphasizes the integration of mature digital monitoring solutions into a coherent system of systems that utilize AI enabled analytics, data-driven innovations, and user-centric designs. By mapping current SHM tools and clarifying their functional roles, this paper contributes to SMARTIN's broader effort to build a shared knowledge base, strengthen evidence informed maintenance practices, and support the scalable uptake of digital solutions across diverse operational contexts.

3 Methodology

This study examines digital solutions for SHM, defined as methods, systems, or platforms that employ digital technologies to observe, assess, or anticipate the condition of transport infrastructure assets. The unit of analysis is the digital solution itself rather than specific deployment site or infrastructure component. A set of thirteen representative solutions is selected to reflect the diversity of current practice.

The selected solutions include different asset types, sensing technologies, and levels of maturity, allowing for comparative analysis without aim to provide exhaustive coverage of the field. An overview of the analyzed solutions and their primary characteristics is provided in Table 1.

Each solution is described using a set of descriptors designed to support structured comparison across otherwise heterogeneous technologies. Specifically, the description includes:

- the types of data employed,
- the dominant analytical approaches,
- the primary functional objective of the solution,
- the spatial scale at which outputs are generated, and
- indicators of technological maturity and validation strength.

In addition, each solution reports key barriers that currently limit wider adoption or transferability.

To ensure comparability, data sources were grouped into modality categories representing common sensing and information streams in infrastructure monitoring, including visual or imaging data, embedded or field-based sensors, active remote sensing techniques, and contextual or environmental information.

Analytical approaches were similarly grouped into high-level categories, such as computer vision, signal processing, classical machine learning, and predictive or forecasting methods. Functional objectives were classified according to whether the solution primarily supports monitoring, prediction, control, or decision support. Spatial scale was defined as the dominant spatial unit at which results are produced, ranging from point or sensor level to asset, segment, corridor, or network scale. Technology Readiness Level (TRL) and evidence quality were used to represent maturity and confidence in reported performance, respectively.

Hierarchical grouping is used as an exploratory tool to illustrate similarities and differences between solutions across the selected descriptors, without imposing predefined categories or fixed classifications. Throughout the analysis, emphasis is placed on identifying similarities among solutions to reveal dominant patterns, functional groups, and recurring gaps that characterize the current state of digital SHM. By structuring heterogeneous solutions within a common descriptive and comparative framework, the methodology supports meaningful synthesis for infrastructure managers, practitioners, and researchers, while aligning with SMARTIN's broader goal of enabling knowledge harmonization, capacity building, and evidence-informed decision making.

Table 1. Overview of digital SHM solutions

Tool	Description
ARTeMIS Modal	A vibration-based SHM software that processes periodic weak-motion recordings and integrates easily with commercial monitoring systems
Active sensor NDT	Ultrasonic wave-based non-destructive testing methods (Impact Echo (IE), Ultrasonic Body Wave (UBW), and Ultrasonic Surface Wave (USW)) used to measure material properties and detect internal defects in concrete and other structures
YOLO-RD	A YOLOv8-based road-damage detector optimized for small, irregular defects using Star Operation Module (SOM), Multi-dimensional Auxiliary Fusion (MAF), and Wavelet Transform Convolution (WTC) modules for improved accuracy
Proqio platform	A unified data-intelligence platform that centralizes, visualizes, and manages monitoring, surveying, and environmental data for infrastructure projects. It allows real-time visualization, alerts, and reporting
GPR-NDT system	A combined Ground-Penetrating Radar (GPR) and Non-Destructive Testing (NDT) approach for surface and subsurface pavement assessment, supporting preventive maintenance and integration with GIS, Building Information Modeling (BIM), and Digital Twins for visualization and real-time monitoring
BIM Digital twin SHM	A BIM-based digital twin system that integrates multi-sensor bridge data with 3D visualization and cloud analytics for real-time monitoring and early warnings
Triple-threshold crack detector	An image-based pavement crack detection method using a structured forest classifier and triple-threshold processing to enhance crack continuity and reduce noise
SHM-WIM system	A cloud-based bridge monitoring system combining structural sensors with weigh-in-motion (WIM) data for continuous condition assessment and overload detection
Autonomous UAV-SHM	A GPS-independent Unmanned Aerial Vehicle (UAV) structural inspection system using ultrasonic beacons localization and deep convolutional neural networks (CNNs) for automated visual damage detection and geo-tagging
Smartphone road damage detector	A CNN-based system that detects and classifies road surface defects from smartphone images captured during vehicle travel

(*continued*)

Table 1. (*continued*)

Tool	Description
OMICRON Inspection	A suite of automated road-inspection solutions including multi-UAV systems, long-range UAVs, V2I/I2V road communication, and advanced sensor combinations aimed at improving safety, reducing traffic disruption, and automating inspections
OMICRON Predictive	A predictive-maintenance platform integrating inspection data into a road digital twin and a decision support system for condition forecasting and intervention planning
OMICRON Bridge	A modular hybrid-bridge retrofitting system paired with a BIM/GIS digital twin for lifecycle monitoring to reduce disruption and maintenance risk

4 Results

4.1 Data Modalities and Analytical Approaches

The thirteen SHM solutions analyzed exhibit a high degree of technological orientation. This is also true for the analytical procedures used. Computer vision methods and defect detection algorithms dominate the landscape, primarily supporting the identification of surface-level damage, cracks, and anomalies in pavements and structural elements. Signal processing and vibration-based analysis methods form a second large group. These are mostly used in SHM applications where time-series sensor data are analyzed to detect changes in dynamic or modal properties. Machine learning techniques, such as tree-based classifiers, appear across both vision-based and sensor-based solutions, while predictive and forecasting methods are mainly associated with more mature monitoring and maintenance platforms.

Overall, it appears that currently, SHM field is driven primarily by data-intensive detection and condition assessments, which emphasize perception and pattern recognition over an integrated prediction or intervention approach. The set of digital solutions included in the analysis, together with their primary characteristics, is summarized in Table 2.

Table 2. SHM solutions with their key characteristics

Solution	Infrastructure type	Primary data source	Main function
ARTeMIS Modal (Pro/SHM)	Bridges, civil structures	Vibration sensors	Structural monitoring
Active sensors for health monitoring	Bridges, structures	Embedded sensors	Structural monitoring
YOLO-RD	Roads, pavements	Images/video	Damage detection

(*continued*)

Table 2. (*continued*)

Solution	Infrastructure type	Primary data source	Main function
proqio	Roads, urban assets	Visual data, mobile sensors	Condition assessment
GPR + NDT combined approach	Roads, bridges	GPR, NDT sensors	Subsurface inspection
BIM-based digital twin SHM	Buildings, bridges	Sensor data, BIM models	Monitoring & decision support
Triple-Threshold Pavement Crack Detection (RSF)	Roads, pavements	Images	Crack detection
SHM combined with Weigh-in-Motion (WIM)	Bridges	Vibration, WIM data	Monitoring & degradation assessment
UAV-based SHM with deep learning	Bridges, structures	UAV imagery, ultrasonic sensors	Damage detection
Smartphone-based road damage detection (DNN)	Roads	Smartphone images	Damage detection
OMICRON inspection solutions	Power & transport assets	Sensor & inspection data	Asset inspection
OMICRON predictive maintenance solutions	Power & transport assets	Sensor & operational data	Predictive maintenance
OMICRON bridge retrofitting solutions	Bridges	Inspection & structural data	Decision support

4.2 Spatial Granularity and Maturity Characteristics

Most of the sample solutions operate at the asset or segment level, focusing on individual bridges, pavement sections, or structural components. An "asset" represents a discrete, manageable unit of infrastructure whose condition can be independently monitored and maintained. This spatial focus represents the reality of how infrastructure inspections and maintenance occur, i.e., localized condition information is needed in order to plan and prioritize interventions. Table 3 supports this observation by illustrating the application location and scale each solution was applied at. It shows also that the majority of the solutions were developed at the asset level across both emerging and matured tools.

TRL varies among all the solutions studied. Commercial inspection and predictive maintenance platforms generally possess greater levels of maturity than non-commercial solutions, as they have demonstrated their ability to perform in actual operating environments. Conversely, most research-based methods, particularly those relying on advanced computer vision or hybrid sensing approaches, demonstrate relatively low TRL. Generally, this occurs because these types of tools have either limited demonstration in the

field, rely heavily on curated data sets, and/or are sensitive to site-specific conditions. These results also suggest that there remains a persistent gap between new methodologies and their successful integration into large-scale operational practices for SHM.

Table 3. Deployment sites and scales of application for each tool

Tool	Deployment	Scale	TRL
ARTeMIS Modal	Operational deployments and field tests	Asset	9
Active sensor NDT	Experimental and pilot studies	Asset	8
YOLO-RD	Dataset-based evaluation	Segment	5
Proqio platform	Pilot deployments	Asset/Segment	9
GPR-NDT system	Research case studies	Segment	5
BIM Digital twin SHM	Pilot implementation	Asset	8
Triple-threshold crack detector	Dataset-based evaluation	Segment	5
SHM-WIM system	Operational deployments	Asset	9
Autonomous UAV-SHM	Experimental testing	Asset	4
Smartphone road damage detector	Large-scale pilot deployment	Segment/Network	7
OMICRON Inspection	Operational deployment	Asset	6–7
OMICRON Predictive	Operational deployment	Asset/Network	7
OMICRON Bridge	Decision-support application	Asset	7

4.3 Grouping of SHM Solutions

The solutions show coherent grouping based upon common technical and functional features. Three groups are identified: a) Vision-based and sensor-driven monitoring tools for the purpose of detection and condition assessment, b) System-oriented SHM & predictive maintenance, and c) Specialized subsurface & intervention solutions.

Solutions within the first group share similar data modalities, analytical approaches, and functional objectives. Solutions in the second group place greater emphasis on time-series analysis, forecasting, and integration with asset management workflows, and they are generally associated with higher maturity levels and stronger evidence bases. Subsurface inspection approaches, in the third group, appear more distinct due to their specialized sensing requirements and analytical processes. Table 4 presents the mapping of SHM solutions to their primary data modalities.

Table 4. SHM solutions and data modalities

Data modalities	Active sensor NDT	YOLO-RD	Proqio platform	Triple-threshold crack detector	Autonomous UAV-SHM	Smartphone road damage detector	ARTeMIS Modal	BIM Digital twin SHM	SHM-WIM system	OMICRON Inspection	OMICRON Predictive	GPR-NDT system	OMICRON Bridge
	Vision-based & sensor-driven monitoring						System-oriented SHM & predictive maintenance					Specialized subsurface & intervention solutions	
Active remote sensing data										✓	✓	✓	
Digital and network-based data feeds				✓	✓				✓		✓		✓
Environmental and contextual data				✓					✓			✓	
Historical and archival data												✓	
Infrastructure-embedded and field sensors			✓	✓			✓		✓			✓	✓
Positioning and motion sensors	✓		✓		✓	✓				✓	✓	✓	
Visual and imaging sensors		✓			✓	✓		✓		✓	✓	✓	✓

The tools are also grouped upon the type and availability of data on which they rely. As shown in Table 4, vision-based and sensor-driven monitoring tools tend to be dependent on large volumes of labeled or high-frequency data. This reinforces their similarity in both analytical approaches and functional scopes. In contrast, system-oriented and predictive solutions are mainly supported by longer-term time-series data and operational records. Therefore, they would integrate well with asset management workflows, while imposing higher data consistency and quality requirements. Specialized subsurface inspection approaches form a more distinct group due to reliance on customized or site-specific datasets.

4.4 Reported Gaps and Deployment Barriers

Building on the comparative analysis of solution groupings, this section identifies key gaps and barriers that limit the deployment, transferability, and integration of digital SHM solutions.

Vision-based and sensor-driven monitoring solutions appear to face challenges related to data availability and transferability. These tools typically rely on large volumes of labeled visual or sensor data for training and validation; however, such datasets are often scarce, heterogeneous, or context-specific. As a result, performance and reliability can vary significantly across locations, asset types, and environmental conditions. Additionally, many of these solutions remain weakly integrated with asset management or decision-support systems, limiting their ability to use detection results for improving maintenance planning.

System-oriented SHM and predictive maintenance solutions have their own unique, yet similarly enduring barriers. These solutions can be much more mature and typically

well-integrated into operational workflows. However, they rely upon the existence of long-term, high-quality time-series data from multiple sources. The primary obstacle that is experienced by most users of these solutions is complexity associated with integrating various disparate data sources, particularly when combining heterogeneous sensor streams, legacy databases, and operational records. A major impediment to widespread use of these solutions are issues of interoperability, standardization, and sustaining the quality of the data. In several cases, high deployment and maintenance costs are identified as an additional obstacle to wider adoption, especially for resource-constrained infrastructure operators.

Specialized subsurface inspection and intervention solutions, such as ground-penetrating radar–based approaches and bridge retrofitting support tools, appear to have a clearer distinction in the similarity landscape, and are facing fewer but more severe challenges. The primary reasons these solutions can be difficult to scale up and apply to other sites is due to customized sensor configurations and site specific calibrations. Additionally, high complexity and operating cost, along with poor integration with broader monitoring and management frameworks, further limit their use beyond targeted inspection or interventions.

All three groups demonstrate similar gap areas related to needs for improved data standardization, better integration of heterogeneous data sources, and clearer pathways from monitoring outputs to decision-support and control mechanisms. These findings emphasize the need for knowledge-based platforms, such as those being envisioned through the SMARTIN project, that identify both technological capabilities and recurring barriers and contextual constraints. By aggregating the gap areas by solution group, the analysis provides an opportunity for future research, standardization efforts, and capacity-building actions relevant to infrastructure managers and policy makers.

To synthesize these observations and make the comparison across solution groups explicit, Table 5 consolidates the reported gaps and deployment barriers identified in the reviewed solutions. The table aligns individual tools and platforms with recurring technical, operational, and institutional challenges, highlighting both group-specific and cross-cutting limitations. It illustrates how vision-based and sensor-driven approaches are predominantly constrained by data-related issues (e.g., labeled datasets, multimodal integration, and real-time data streams), while system-oriented SHM and predictive maintenance solutions are more frequently affected by deployment costs. At the same time, specialized subsurface and intervention-oriented solutions stand out for their site specificity, and need of data.

Table 5. Key barriers of SHM solutions

Key gaps	Active sensor NDT	YOLO-RD	Proqio platform	Triple-threshold crack detector	Autonomous UAV-SHM	Smartphone road damage detector	ARTeMIS Modal	BIM Digital twin SHM	SHM-WIM system	OMICRON Inspection	OMICRON Predictive	GPR-NDT system	OMICRON Bridge
	Vision-based & sensor-driven monitoring						System-oriented SHM & predictive maintenance					Specialized subsurface & intervention solutions	
High data processing and communication demands								✓					
Highly site- or mode-specific (difficult to generalise)									✓				✓
Insufficient real-time data streams		✓		✓	✓								
Lack of economic viability for wide deployment											✓		
Lack of high-quality labelled datasets		✓				✓						✓	
Lack of multimodal data integration (e.g. GNSS + video + IoT)	✓		✓	✓								✓	
Low TRL/SRL (immature technology or social adoption)					✓					✓	✓		
Missing guidelines or best practices for adaptation	✓										✓		
Requires labelled datasets for training				✓	✓								
Requires proprietary or costly hardware/ software	✓		✓				✓	✓	✓	✓			✓

5 Discussion

The results of this study provide a structured mapping of a sample of existing digital SHM solutions that offer insights directly relevant to ongoing efforts in knowledge harmonization and evidence-informed infrastructure management. When diverse tools are fitted into a shared explanatory framework, the interpretation can follow systematically through technological innovations and changes that may be otherwise difficult to compare and contrast because of differences in scale, development stage and terminologies.

One of the most notable conclusions is the very large concentration of solutions dedicated to monitoring and detection. Vision-based inspection systems and sensor-based SHM systems still represent the majority of existing solutions, accounting for the most recent technologies and the highest complexity levels in the field, given the maturity of data acquisition technologies. However, they are a few steps away from condition assessment functionalities towards predictive or control functionalities. Instead, this bias represents a dis-junction in the production of condition data and its application to actionable asset management processes, such as planning interventions and so on. Bridging this gap is critical to moving reactive infrastructure management paradigms forward to proactive infrastructure management paradigms.

The categorization of solutions based on sensing modalities and analytical approaches also stresses the fragmented nature of the current spectrum. While these categories may be intuitive from a technical standpoint, they also demonstrate some limitations in cross-type interoperability and integration. The three groups co-evolve primarily in parallel with little means for merging complementary data sources or analytical results across the stages of monitoring, predicting, and acting. Understand the system-level, this fragmentation limits the potentials of digital solutions in supporting full infrastructure assessments and integrated management strategies and thus calls for integrative frameworks that can connect diverse monitoring.

The examination of identified gaps and obstacles suggests that numerous challenges in digital SHM are systemic rather than pertaining to specific solutions. Some of the problems that arise are that there are not enough labeled datasets, that it is hard to combine data from different sources, and that proprietary hardware or software components make things harder. These problems show up in different types of solutions and at different stages of maturity, and they match what has been reported in the literature. Their continued existence underscores the necessity for synchronized initiatives in standardization, data-sharing protocols, and methodological clarity. The analysis organizes and groups these barriers in a way that makes it easier to decide the order of research, development, and standardization activities.

6 Conclusions

The study has undertaken a structured comparative analysis of digital solutions for SHM using a set of thirteen representative methods and tools. The organization of heterogenous digital solutions into a structured framework allows for an effective and structured comparison of tools that are generally evaluated in isolation. The analysis highlights dominant groups of technologies, differences in maturity and validation, and common barriers to wider deployment.

The findings reveal a strong concentration of current practice on monitoring and detection functions, with vision-based and sensor-driven systems dominating the field. The results also highlight a significant lack of digital solutions aimed at predictive analytics, decision support or controlling functionalities which shows a gap between condition assessments and actions taken to manage the performance of infrastructure. The grouping of solutions into three categories reveals a fragmented digital SHM ecosystem. There are limited opportunities for interoperability and poor integration of complementary data sources which prevents the development of holistic, system-wide management approaches. Therefore, there is a need for integrative frameworks that can effectively link monitoring, prediction and intervention.

The analysis of identified gaps indicates that many barriers are systemic, such as insufficient real-time data streams, problems associated with multimodal data integration, and reliance on proprietary or costly technologies. By documenting these issues in a structured format, this study provides a basis for the prioritization of future research and standardization initiatives.

Finally, the results show how comparative analysis can support decision-making by infrastructure managers, public authorities, and mobility practitioners. By clarifying relationships between solutions, identifying areas of overlap and complementarity,

and highlighting persistent gaps, the analysis supports informed technology selection, benchmarking, and integration. In this way, the study contributes both to academic synthesis and to practical efforts aimed at supporting evidence-based adoption of digital solutions for SHM.

Acknowledgments. This work was funded by the European Union's Horizon 2020 research and innovation programme through the project SMARTIN (Grant Agreement No. 101203240). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Domaneschi, M., Cucuzza, R., Martinelli, L., Noori, M., Marano, G.C.: A probabilistic framework for the resilience assessment of transport infrastructure systems via structural health monitoring and control based on a cost function approach. Struct. Infrastruct. Eng. 1–13 (2024). https://doi.org/10.1080/15732479.2024.2318231
2. Alonso-Díaz, A., Solla, M., Rodríguez, J.L., Elseicy, A.: Integrated approach to improve road infrastructure resilience through complementary non-destructive techniques. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. **X-4/W5-2024**, 9–15 (2024). https://doi.org/10.5194/isprs-annals-X-4-W5-2024-9-2024
3. Li, X., Dong, Y.-X., Xiang, W.: Developing a BIM based digital twin system for structural health monitoring of civil infrastructure. Meas. Sci. Technol. **35**, 115117 (2024). https://doi.org/10.1088/1361-6501/ad730f
4. Fanijo, E., Liu, J., Mohammed, T.: Transportation infrastructure health monitoring. In: Infrastructure Development - Innovations, Challenges and Opportunities [Working Title]. IntechOpen (2025). https://doi.org/10.5772/intechopen.1012341
5. Cavalli, M.C., et al.: Review of advanced road materials, structures, equipment, and detection technologies. J. Road Eng. **3**, 370–468 (2023). https://doi.org/10.1016/j.jreng.2023.12.001
6. Elseicy, A., Alonso-Díaz, A., Solla, M., Rasol, M., Santos-Assunçao, S.: Combined use of GPR and other NDTs for road pavement assessment: an overview. Remote Sens. **14**, 4336 (2022). https://doi.org/10.3390/rs14174336
7. Kim, J.-W., Choi, H.-W., Kim, S.-K., Na, W.S.: Review of image-processing-based technology for structural health monitoring of civil infrastructures. J. Imaging **10**, 93 (2024). https://doi.org/10.3390/jimaging10040093
8. Liu, C., Zhang, P., Xu, X.: Literature review of digital twin technologies for civil infrastructure. J. Infrastruct. Intell. Resilien. **2**, 100050 (2023). https://doi.org/10.1016/j.iintel.2023.100050
9. Mardanshahi, A., Sreekumar, A., Yang, X., Barman, S.K., Chronopoulos, D.: Sensing techniques for structural health monitoring: a state-of-the-art review on performance criteria and new-generation technologies. Sensors **25**, 1424 (2025). https://doi.org/10.3390/s25051424
10. Sakr, M., Sadhu, A.: Recent progress and future outlook of digital twins in structural health monitoring of civil infrastructure. Smart Mater. Struct. **33**, 033001 (2024). https://doi.org/10.1088/1361-665X/ad2bd7
11. Wu, S., et al.: Evaluation of smart infrastructure systems and novel UV-oriented solution for integration, resilience, inclusiveness, and sustainability. In: 2020 5th International Conference on Universal Village (UV), pp. 1–45. IEEE, Boston, MA, USA (2020). https://doi.org/10.1109/UV50937.2020.9426194

12. Alqasi, M.A.Y., Alkelanie, Y.A.M., Alnagrat, A.J.A.: Intelligent infrastructure for urban transportation: the role of artificial intelligence in predictive maintenance. Brilliance **4**, 625–637 (2024). https://doi.org/10.47709/brilliance.v4i2.4889
13. Bango, M., Mkandawire, M.: Predictive maintenance in bridge infrastructure. IJARSCT 240–243 (2024). https://doi.org/10.48175/IJARSCT-18235
14. Plevris, V., Papazafeiropoulos, G.: AI in structural health monitoring for infrastructure maintenance and safety. Infrastructures **9**, 225 (2024). https://doi.org/10.3390/infrastructures9120225
15. Elsisi, A., Zamrawi, A., Emad, S.: A comprehensive review of structural health monitoring for steel bridges: technologies, data analytics, and future directions (2025). https://www.preprints.org/manuscript/202510.1575/v1. https://doi.org/10.20944/preprints202510.1575.v1
16. Laflamme, S., et al.: Roadmap on measurement technologies for next generation structural health monitoring systems. Meas. Sci. Technol. **34**, 093001 (2023). https://doi.org/10.1088/1361-6501/acd135
17. Sause, M.G.R., Jasiūnienė, E. (eds.): Structural Health Monitoring Damage Detection Systems for Aerospace. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72192-3
18. Sebastian, R., Weise, M., Mitsch, N., Giurgiu, I., Rodríguez, A.S.: Towards European standardization of digitalization approaches for monitoring and safety of bridges and tunnels. In: ECPPM 2022 - eWork and eBusiness in Architecture, Engineering and Construction 2022, pp. 204–210. CRC Press, London (2023). https://doi.org/10.1201/9781003354222-26
19. Gharehbaghi, V.R., et al.: A critical review on structural health monitoring: definitions, methods, and perspectives. Arch. Comput. Meth. Eng. **29**, 2209–2235 (2022). https://doi.org/10.1007/s11831-021-09665-9
20. Cha, Y.-J., Ali, R., Lewis, J., Büyüköztürk, O.: Deep learning-based structural health monitoring. Autom. Constr. **161**, 105328 (2024). https://doi.org/10.1016/j.autcon.2024.105328
21. Firoozi, A.A., Firoozi, A.A.: Prognostic methods in structural health monitoring of large-scale structures: a PRISMA-guided systematic review (2024). https://www.ssrn.com/abstract=4702497, https://doi.org/10.2139/ssrn.4702497
22. Numan, M.: Advancements in structural health monitoring: a review of machine learning approaches for damage detection and assessment. IJCCSE **20**, 124–142 (2024). https://doi.org/10.22337/2587-9618-2024-20-1-124-142
23. Hassani, S., Dackermann, U., Mousavi, M., Li, J.: A systematic review of data fusion techniques for optimized structural health monitoring. Inform. Fus. **103**, 102136 (2024). https://doi.org/10.1016/j.inffus.2023.102136
24. Azad, M.M., Kim, S., Cheon, Y.B., Kim, H.S.: Intelligent structural health monitoring of composite structures using machine learning, deep learning, and transfer learning: a review. Adv. Compos. Mater **33**, 162–188 (2024). https://doi.org/10.1080/09243046.2023.2215474
25. Wang, H., Barone, G., Smith, A.: Current and future role of data fusion and machine learning in infrastructure health monitoring. Struct. Infrastruct. Eng. **20**, 1853–1882 (2024). https://doi.org/10.1080/15732479.2023.2165118
26. Jia, J., Li, Y.: Deep learning for structural health monitoring: data, algorithms, applications, challenges, and trends. Sensors **23**, 8824 (2023). https://doi.org/10.3390/s23218824
27. Eltouny, K., Gomaa, M., Liang, X.: Unsupervised learning methods for data-driven vibration-based structural health monitoring: a review. Sensors **23**, 3290 (2023). https://doi.org/10.3390/s23063290
28. Lawal, O., V.Shajihan, S.A., Mechitov, K., Spencer Jr., B.F.: Edge integration of artificial intelligence into wireless smart sensor platforms for railroad bridge impact detection (2024). https://www.preprints.org/manuscript/202408.0862/v1, https://doi.org/10.20944/preprints202408.0862.v1

29. Mondal, T.G., Chen, G.: Artificial intelligence in civil infrastructure health monitoring—historical perspectives, current trends, and future visions. Front. Built Environ. **8**, 1007886 (2022). https://doi.org/10.3389/fbuil.2022.1007886
30. Rezazadeh, N., De Luca, A., Perfetto, D., Salami, M.R., Lamanna, G.: Systematic critical review of structural health monitoring under environmental and operational variability: approaches for baseline compensation, adaptation, and reference-free techniques. Smart Mater. Struct. **34**, 073001 (2025). https://doi.org/10.1088/1361-665X/ade7db
31. Preethichandra, D.M.G., Suntharavadivel, T.G., Kalutara, P., Piyathilaka, L., Izhar, U.: Influence of smart sensors on structural health monitoring systems and future asset management practices. Sensors **23**, 8279 (2023). https://doi.org/10.3390/s23198279
32. Anastasopoulos, M., et al.: Design, development and field trial evaluation of an intelligent asset management system for railway networks. In: Kumar, U., Karim, R., Galar, D., Kour, R. (eds.) International Congress and Workshop on Industrial AI and eMaintenance 2023, pp. 279–289. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-39619-9_20
33. Liew, J., Rashidi, M., Le, K., Nazar, A.M., Sorooshnia, E.: Reviewing a decade of structural health monitoring in footbridges: advances, challenges, and future directions. Remote Sens. **17**, 2807 (2025). https://doi.org/10.3390/rs17162807
34. Zanelli, F., Debattisti, N., Mauri, M., Argentino, A., Belloli, M.: Development and field validation of wireless sensors for railway bridge modal identification. Appl. Sci. **13**, 3620 (2023). https://doi.org/10.3390/app13063620

Applying Transdisciplinary Experimental Methods to the Regeneration of a Former Nunnery

Teresa Pedretti[1](✉), Pietro V. Ambrosini[2], Alessandro Busana[2], Daniele Cappelletti[2], and Letizia Bollini[1]

[1] Free University of Bolzano-Bozen, Bolzano, Italy
teresa.pedretti@student.unibz.it
[2] Campomarzio Collective, Bolzano-Trento, Italy

Abstract. In the fields of architecture and urban planning, research over recent decades has shifted its focus from expansion to the transformation and regeneration of existing urban structures and buildings. In this context, scholars have primarily sought tools and methods to optimize building renovation processes and enhance their energy efficiency and structural performance. This article considers regeneration as more than a technical issue and proposes an alternative strategy to integrate urban analysis, building analysis, stakeholder mapping, placemaking, and scenario building into a unified process including people from the very beginning too. The contribution, therefore, through the description of a case study involving the development of a feasibility study for the regeneration of a former nunnery in Italy, illustrates a three-stage framework: 1) analysis, 2) discovery, 3) valorization, guided by a research-through-design methodology. It presents a vision of regeneration as a space where design can assume innovative roles, fostering community involvement in shaping urban transformation strategies. Most importantly, it can serve as a means to reconfigure disciplinary knowledge into transdisciplinary frameworks, addressing questions related to the future regeneration of heritage assets with uncertain uses.

Keywords: urban regeneration · transdisciplinary planning · participatory design

1 Introduction and Context

1.1 The Sense of Regeneration in the Neoliberal Era

The conclusion of the modern era, coupled with the decommissioning of industrial zones, the relocation of manufacturing activities, and the globalization of markets, has resulted in significant portions of the city remaining unused and awaiting repurposing to serve the needs of a transitioning society [1, 2].

Since the 1990s, in Italy, as well as across Europe and much of the United States, the discourse surrounding urban development has progressively shifted from an expansion-oriented research orientation to a consideration of the transformation of existing urban

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 346–361, 2027.
https://doi.org/10.1007/978-3-032-30494-0_22

structures [3]. In this context of debate and the difficulty of establishing new regulatory schemes, many cities have experienced market-driven urban transformations grounded in neoliberal principles, operating under the assumption that actors would self-regulate [4–6].

The acknowledgment of the market's inability to self-regulate, coupled with protests spreading across numerous cities and the extensive impacts of urban transformation driven by competition to attract increasing capital, has in recent years led to the dissemination of various practices and tools. Additionally, there has been a renewed emphasis in academic literature on experimenting with urban planning models that prioritize the social and environmental sustainability of urban transformations [7–9].

The dissemination of gentrification [10–12], the substitution or mixing of the historical inhabitants of an area with new inhabitants capable of purchasing properties made expensive by their transformation [13], the expulsion of the less well-off classes from redeveloped urban areas from Milan [14, 15] Barcelona [12] and Belgrade [16]; the commodification of historical centres and the worsening of overtourism [17] are phenomena that contribute to the development of new sensibilities and the desire to experiment with new ways of imagining, designing and realising different forms of urban transformation.

The diffusion of practices of community involvement is a line of actions and experiments that, born in the Sixties [18, 19], has regained vigor [20] after a couple of decades of stagnation. Scholars in urban planning, sociology, and political science developed research frameworks, theories, and perspectives focused primarily on techniques and tools for collaborative, communicative, and radical planning [21, 22], as well as placemaking [23] and democratic experimentation [24, 25].

The last two decades have also seen growing interest in citizen involvement in urban transformations, particularly within design disciplines such as participatory design, civic design, and policy design [26]. More specifically, participatory design [27] has been engaged in searching for tools and methods to plan and run workshops and actions, improve citizen engagement, enhance the accessibility of these urban laboratories [28], and understand how, during the process, the public gathers around emerging cities' issues [29–31].

The method tested through the case study illustrated below is part of a line of research that goes beyond disciplinary contexts to develop tailor-made models and methodological approaches, drawing on the potential of a perceived all-encompassing unity of knowledge [32]. Many scholars indeed concentrate their research on integrating participatory methods as tools for developing disciplinary knowledge; however, less emphasis has been placed on exploring new theoretical frameworks that enable these assemblies not only to become collective, shared knowledge but also to serve as a place where transdisciplinary research through design in the field of urban planning and urban design could grow.

This paper falls within the scope of the research-through-design field [33, 34]. Therefore, it develops a methodological proposal following the completion of a participatory regeneration project, or, rather, a theoretical proposal resulting from the methodological knowledge generated by the project [35].

1.2 The Context

Casa Sant'Angela is a structure on Via Rosmini in Trento, originally built to accommodate the Ursuline Sisters' congregation. Erected in the early 1930s, it is positioned at the periphery of the historic center, in a zone that, over the past two centuries, has been swiftly integrated into the urban fabric of the contemporary city, evolving with a diverse morphological configuration over time. The complex was constructed on a residual lot that, until the 19th century, was situated in open countryside. Its development was impacted after World War II by the construction of the Brenner railway, which confined it to a strip of buildings between the road and the tracks, rendering it a peripheral area effectively separated from the social activity associated with other social hubs.

The Casa Sant'Angela complex, however, now forms an integral part of a rapidly evolving urban landscape. Previously situated on the periphery and prominently defined by the presence of the Monumental Cemetery, this area is progressively adopting a new function as a nexus of urban flows, activities, and functions.

The gradual abandonment of the facility by the Sisters of the Company of Saint Ursula and the outbreak of the Covid-19 pandemic ultimately resulted in the repurposing of the complex for emergency functions. This included the establishment of accommodation and the daytime distribution of meals to individuals in need. Such temporary utilization, through practical application, demonstrated the new potential of these spaces, thereby opening the complex to a different user demographic and fostering the development of a newly hosted community.

Based on these encouraging temporary uses, the Management Board of the Trento European Capital of Volunteering Committee commissioned a multidisciplinary team to conduct a feasibility study regarding the regeneration and transformation of Casa Sant'Angela into a community space.

1.3 Experimental Method and Approaches Design

While organizing the actions needed to develop the feasibility study, it was decided to consider the community of local third sector organizations, social enterprises, and cultural enterprises as a key element. Therefore, it was determined to give an important role to the design of participatory processes [27, 36, 37], alongside the survey and analysis of the building's structural, architectural, and urban features. To achieve this scope, the tested methodology borrowed from the approach to social innovation developed in design disciplines [38] and in studies on the relationship between design and territory [39].

The process based on these foundations consists of three main stages: 1) analysis; 2) discovery; 3) valorization.

The analysis stage (1) is fundamental to experimentation. It is inspired by the holistic territorial survey paradigm, placing the community at the center and shifting focus from just solving a problem and achieving a result to reflecting on the reasons behind the intervention. By developing a framework that considers the building's urban and architectural evolution, existing structural and functional constraints, and reconstructing the complex's functions and role in the city, the analysis phase reveals potential and lays the groundwork for innovative projects [40].

This phase involves the community's contribution to mapping resources for understanding the historical and social constructs within which the heritage is embedded. Subsequently, the community assumes the role of protagonist during the discovery phase (2), where the narrative and sharing of the framework developed during the analysis phase enable participants to "discover" the artifacts involved and the relationships of interest, frequently altering their perceptions of what already exists.

The valorization phase (3) develops from the results of the previous two phases. During this stage, reworking analyses, shifting perspectives on artifacts and their relationships, and engaging in open, shared reflection enable the reconfiguration of existing resources. This process supports the community's shared social goals and guides a development project rooted in local reality and its resources.

2 Case Study Casa Sant'Angela

2.1 First Phase: Analysis

Following the planned action, the process to deliver the feasibility study for the regeneration of Casa Sant'Angela began with a thorough analysis that included: 1) the location of the complex in relation to the current and future urban layout of the area; 2) a description of the spaces available into the building in terms understandable to all stakeholders; 3) collaboration with organizations interested in creating a map of the existing structures and exploring potential new relationships within the context.

1. Location of the complex in relation to urban layout

The Casa Sant'Angela complex is situated within a rapidly evolving urban setting. It is situated at the intersection of two major roads, increasingly central to the area's accessibility and growth. Along Via Rosmini and Via Giusti, respectively to the north and south of the building, are significant educational institutions: on one side, the Department of Humanities and Philosophy at the University of Trento, and on the other, two higher education schools. These institutions draw thousands of students daily, many of whom travel from outside the province.

The ancient Renaissance route of Via Madruzzo, recently improved by the construction of a pedestrian underpass connecting the city center to the museum complex comprising the MUSE and the Palazzo delle Albere, runs transversely to these main axes.

Along the westward extension of the Adige River, many redevelopment projects are in progress on abandoned structures. Although the early phase has been somewhat uneven, these efforts signal a significant renewal of the cityscape. Notable projects include restoring the Barchesse of the Palazzo delle Albere, converting the old trade fair zone, and establishing a center for cultural and entrepreneurial innovation, currently housing the Santa Chiara Cultural Center.

This urban transformation plan is further strengthened by plans to bury the railway corridor and redevelop the area of the Briamasco Stadium. Viale Madruzzo will not only connect the historic center to the Albere district but also become the center of a new urban network that extends uniquely from the city center to the banks of the Adige River (Fig. 1).

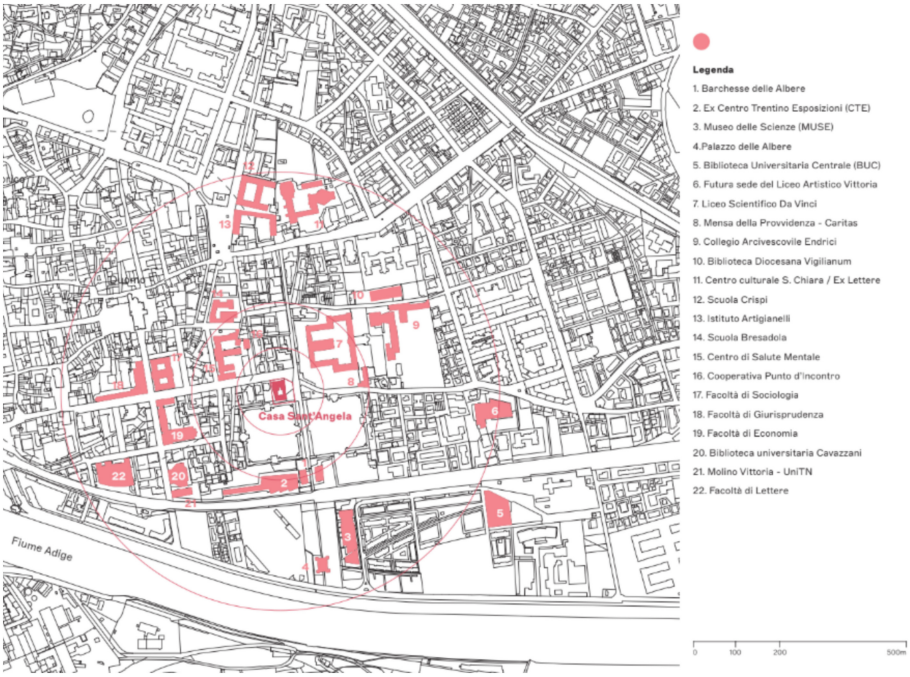


Fig. 1. Casa Sant'Angela in the urban context

2. Description of the spaces in the building

The Compagnia di Sant'Orsola complex results from combining several buildings, each simple and identified by its architectural design. Together, these volumes create a maze-like layout that feels more like a home than a public space. Despite its large size and expansive surfaces, a careful look shows that the complex functions as a large, shared residence.

The residential areas are located on the first and second floors, while the shared spaces, related to logistics and community activities, are mainly on the ground floor and in designated sections of the building.

Considering this peculiarity, reflected both in the material structure of the spaces and their interconnectedness, it was suitable to analyze the complex by breaking it into building blocks, thus clearly emphasizing its constraints and potential for change.

The outdoor spaces are probably the least prioritized resource in the entire complex, even though they cover more than half of the total area. Long kept out of public access, they have stayed in the background, despite being a key part of the site's layout and future potential. Over time, these areas have mainly served functional purposes related to the buildings, such as access and service routes, or have gained decorative value, such as the garden surrounding the entrance on Via Rosmini.

3. Map of the existing and potential new relationships

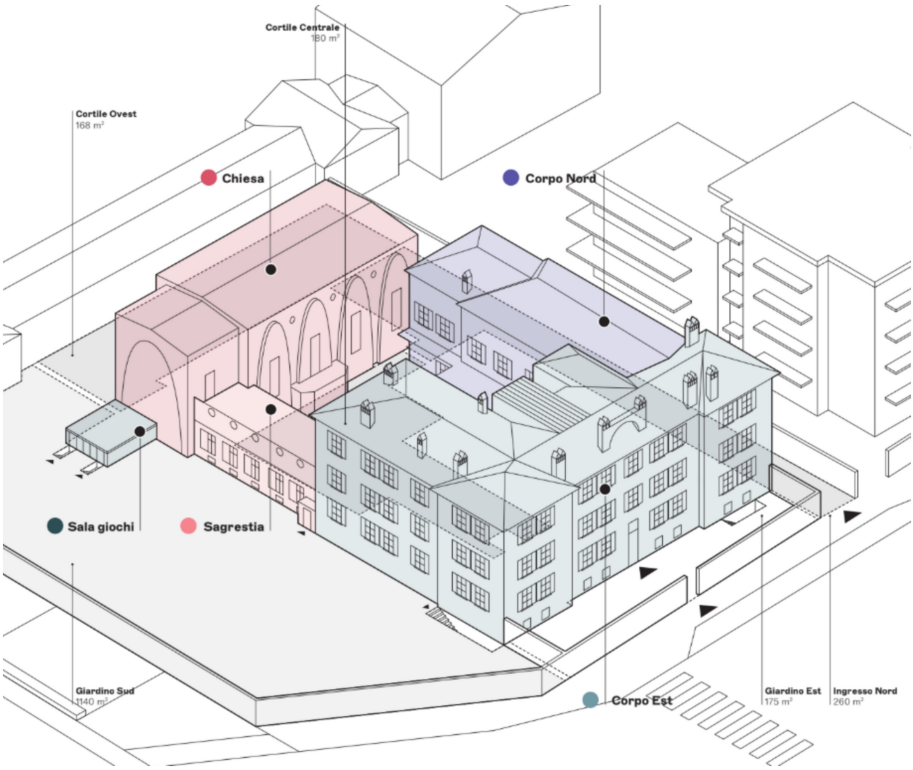


Fig. 2. Description of the building's blocks

We can see buildings in a city not only as places where people arrive, meet, and go daily, but also as hubs that both influence and are influenced by the network they are part of.

Enhancing Casa Sant'Angela's capacity to serve as a hub that fosters connections with the surrounding community is essential today. This ensures that formal and informal networks of people, associations, and institutions can create opportunities for collaboration and innovation.

To understand what kind of hub this building could become from the perspective of the urban network, in collaboration with Trento's social and cultural organizations, a map was created that connects the cultural and social aspects dedicated to the community. This map highlights the spaces, organizations, and stakeholders currently active. For each of these places, Casa Sant'Angela could serve as a development laboratory, a venue for events, or a partner to expand the network of activities. Similarly, these spaces are crucial for Casa Sant'Angela to define its identity and develop a sustainable project over time (Fig. 3).

This first phase led to conceiving and designing the project's structure as a process of creative recombination of existing resources: a) the social capital represented by local profit, nonprofit, public, or voluntary organizations, and b) the historical and real estate

Fig. 3. Map's detail

heritage, consisting of the building, to achieve socially recognized goals. Within the broad range of possible methods and activities, from the most widespread and well-known facilitation or participation actions to co-design groups with defined end users, we chose to act from the perspective of those who want to *make things happen*, given the experimental nature of the work.

In other words, we adopted tools and methods aimed at shaping existing conversations and fostering temporary, open, and flexible dynamics of discussion and conflict about what and how to do it [38, 41] in subsequent meetings with local public and private entities, in the qualitative and quantitative analysis, and in the drafting of the final product.

The activities carried out to achieve these effects have been:

- **Analyze the policy documents** and reports produced by relevant stakeholders during the year in which the process of Trento European Capital of Volunteering has been run.
- Collect answers with an **online open answer questionnaire** to gather insight and experience, delivered to the mailing list of the stakeholders (60 different organizations) mapped by the Management Committee during the European Capital year;
- **Collect and set benchmarks** with other similar common spaces.
- Realize **open interviews** with strategic stakeholders to deepen the map of emerging social needs.

2.2 Second Phase: Discovery

Analyzing expectations, meeting with local stakeholders, and conducting benchmarking research have helped us conceptualize and formalize several intervention scenarios, which are potential clusters of functions, people, and organizations based on socially and culturally grounded regeneration strategies with a strong identity.

These scenarios are a crucial part of the building regeneration process, as they allow for the practical application of concepts and operating models for cultural and social centers and innovative services by comparing them with familiar spaces and locations in nearby areas.

The design of the scenario to be discussed follows a transdisciplinary approach, enabling the creative re-imagining of the disciplines and the possibilities for their combination [42, 43].

Transdisciplinarity allows designers involved in the process to work between, across, and beyond disciplinary borders (Fig. 2), developing custom-made frameworks and methodological approaches to draw on the potentials of a perceived, overarching unity of knowledge [32] (Fig. 4).

Choosing a transdisciplinary approach appear to be the best way to develop urban regeneration scenarios capable of integrating: a) the refurbishment of infrastructure and systems; b) the definition of new functions that reflect the values that emerged from the Volunteering Capital Year process; c) the design of services for different target groups; d) architectural redevelopment; e) the development of construction site management phases; f) the development of the management model for the regenerated building; and g) the design of the placemaking strategy for a hybrid place that the city of Trento will have to learn to understand.

The process of scenario design and analysis has been guided by a rigorous feedback workshop with local third-sector managers and professionals to discuss previously designed scenarios.

The workshop was designed to consider, first, impressions, suggestions, and ideas from every relevant organization active in the local social and cultural sector, and, second, those from institutions involved in processes at various levels with different roles.

Below is a summary of the scenarios developed for comparison and the reasons mapped through the feedback sessions that led to their exclusion:

→ Creation of a sustainable mobility hub, a lack of such a feature in a city like Trento, which is strongly committed to promoting sustainable lifestyles.

The building is not adjacent to any transportation hubs; there are no stations or park-and-ride facilities nearby. This means that a sustainable mobility hub here would not find its natural user base, consisting of commuters who arrive in the city by bus, car, or train and use bicycles or scooters to get around the city center. The building also spans multiple floors, which limits its potential as a bicycle repair shop or parking lot, since moving vehicles between floors would be complicated and challenging.

→ Creation of a coworking space.

Coworking spaces still represent a service that can trigger innovation processes through cross-fertilization and encounters among people working in knowledge-intensive sectors. In Trento, too, several spaces have been designated for this purpose over the past fifteen years. This function underscores that opening a dedicated coworking

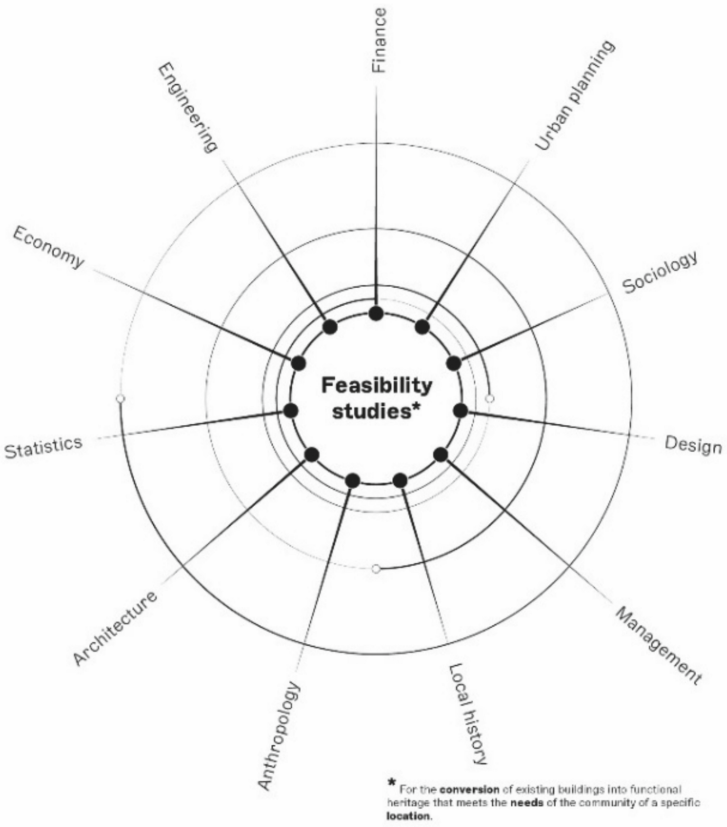


Fig. 4. Disciplines involved in the design of an urban regeneration strategy for a historical building

space within this building would duplicate existing ideas and, therefore, be less likely to spark innovation.

→ Creation of a hub for digital manufacturing, i.e., the broad spectrum of activities related to fablabs and makerspaces, which in recent years has become increasingly central to the design of sustainable forms of urban regeneration. In Trento, there are two such spaces, located at the University and the Muse museum. The possibility of designing new spaces to foster small digital artisan businesses could certainly be a valuable driver of innovation for the city. However, these initiatives require significant mobilization of knowledge and skills from interested parties to launch and sustain in the long term. In other words, spaces of this type could function, upon request, for example, by groups of students or alumni who, after using university spaces, embark on a production initiative.

Among others, two specific strategic scenarios were examined following positive feedback and initial investigations into the current structure.

These are, first, a scenario that transforms Casa Sant'Angela into a cultural center, and second, a scenario that proposes it as a hub for social catering.

The core of a cultural center is its active space, driven by a cultural program structured on two levels: a consistent schedule of cultural activities (such as events, performances, and meetings) and a variety of extracurricular activities, including workshops, one-time events, and rentals for third-party businesses. All these activities, aimed at the public, are managed by multiple entities overseeing the space. It is not necessarily expected that all entities will operate within a formal network or have established a second-level organization.

Social catering is a complex and diverse phenomenon that encompasses all practices that combine food services with social inclusion. Catering services include fine-dining restaurants, osterias and trattorias, pizzerias, and agritourism establishments, as well as itinerant services such as street food, food trucks, home delivery, and catering. Additionally, it can also encompass pastry shops and bakeries on one hand, and bars and pubs on the other. The social catering practices that develop at the intersection of these variables are diverse: from commercial establishments with social missions that employ disadvantaged individuals (possibly reinvesting profits into social projects) to places that offer meals at affordable prices or free of charge to those in need.

The workshop led the process of designing the strategic regeneration scenario into the final phase, in which the multidisciplinary team involved summarized the results of each phase in the final version of the scenario and the feasibility plan. This process results not only in the valorization of the building in sè, but, foremost, in the valorization of the community's contribution and the specificity of the designed strategy.

2.3 Third Phase: Valorization

The main goal of regenerating the Casa Sant'Angela spaces—resulting from participatory processes to define the project and from technical assessments to identify architectural constraints and opportunities—is to develop a broader, more flexible programmatic structure. This aims to create a community space that can accommodate diverse functions and users, spanning various social and cultural spheres.

The main challenge of the strategy is therefore to transform the building into a structure capable of hosting a new community, diverse in its users and dynamic in the recombination of its components. One of the foundational principles of the building's redevelopment is thus the decision to bring all public functions—that is, those open daily to different groups of people—to the ground floor or mezzanine level.

The proposed scenario envisions a clear functional distinction by level, which, with minimal structural work, allows different types of functions to coexist, creating a unified purpose and a center that can define its own identity within the urban landscape.

The mezzanine level is therefore designed to host public activities, while the first and second floors are repurposed for residential use, consistent with the building's original layout and function. Specifically, the mezzanine level features spatial continuity between the different building units, achieved by redefining the interior spaces. While maintaining separate entrances and distinct functions, these spaces can be easily connected, depending on the time of day and the composition of the expected user groups. This flexibility emphasizes the building's public purpose, allowing it to adapt dynamically to the urban environment and community needs by expanding or contracting its multifunctional components.

The upper levels, designated for residential use in this scenario, maintain the division between the two building units, making each floor divisible into two fully autonomous sections. Each level can thus accommodate two separate communities of residents simultaneously.

3 Results

The regeneration strategy for the feasibility plan results in four key components.

1. *The promotion of food as a bridge between cultures and a means of fostering well-being*. The public spaces on the mezzanine levels and those of the church hall define the regeneration strategy, which is grounded in a model that views food, nutrition, and community dining as defining elements for a building that must find its place within the neighbourhood and the city. The functions housed there will be dedicated to promoting healthy eating, health education, culinary instruction and training, and organizing opportunities for gathering and socializing.
2. *Housing to address the housing crisis*. The provision of residential spaces within Casa Sant'Angela addresses the housing needs highlighted by the city of Trento in recent years. Their configuration as separate, non-interconnected spaces does not preclude designing cohabitation projects that bring together diverse communities and groups; rather, it primarily involves arranging the spaces to accommodate the diversity and variety of social needs and emerging vulnerabilities.
3. *Clinics*. Spaces designated as clinics could serve vulnerable individuals to ensure everyone's right to healthcare, or—through a shared commitment to addressing emerging needs—become spaces dedicated to promoting well-being. This could be achieved by preventing new forms of addiction now observed among younger segments of the population, or by providing psychosocial and educational support to families facing social withdrawal.
4. *The Garden of Encounters*. Encountering others, forming new relationships and friendships (but also finding or changing jobs) happen precisely through these kinds of encounters—chance and fortuitous—with people we haven't planned to meet. The garden itself, as a space overlooking the city streets, can serve as the backdrop for these kinds of encounters that create the weak ties essential for improving one's temporary circumstances or personal and professional life.

The final document, elaborated on these key points, outlines:

- the best features to develop to enable efficient and sustainable use of buildings, and the minimum construction and structural interventions necessary to optimize buildings' internal spaces;
- the contextual scenario that makes certain identified functions drive the location's positioning within the geography of the city's public spaces;
- the preliminary estimate of the structure's renovation and operating costs;
- the model for launching the operation of the regenerated space based on a public-private co-designed strategy;
- the business model to ensure the economic and financial sustainability of the management of the regenerated place;

– some final recommendations to avoid the most common mistakes in social innovation and culture-based regeneration processes.

This approach and structure in the final document enable the development of a flexible regeneration strategy focused on the balanced distribution of resources and social capital, tailored to economic sustainability and the actual capacity of local communities to mobilize (Fig. 5).

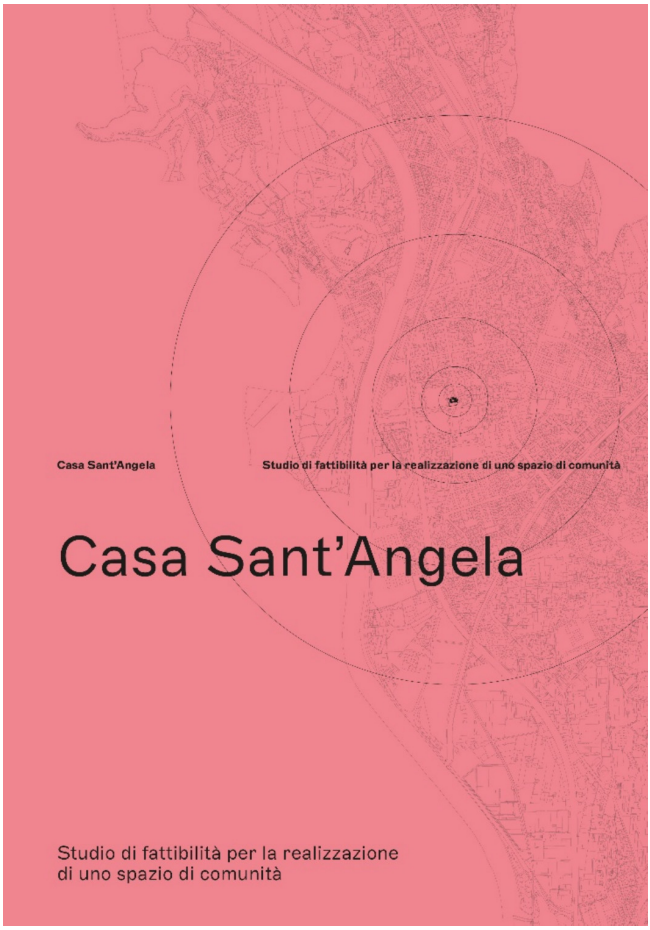


Fig. 5. Final document of the feasibility study for Casa Sant'Angela

4 Conclusion

The experimental approach, first and foremost, required a shift in focus away from a traditional technical consulting model. Whereas, in fact, when faced with the need to renovate a building, clients typically commission an architectural and/or engineering firm, here the mandate went beyond the development of specific solutions aimed at

revitalizing the building. Rather than assuming the role of an external party tasked with organizing the design process into a linear, logical sequence of identifying the elements and requirements necessary for the design solution and developing a solution based on their recombination, our role was therefore to facilitate and activate latent potential.

The focus was therefore directed not so much toward identifying technical and structural solutions to the building's problems to make it usable again by organizations with unknown purposes, but rather toward fostering an understanding of the project context as a complex of architectural, urban, structural, social, and cultural issues whose components are intertwined and often represent conflicting values.

In other words, the approach and method viewed the project context and the issues arising from it as a complex problem [44–46].

This approach led to the adoption of a transdisciplinary approach [42] and the conception of participation as the backbone of the design process. The transdisciplinary approach was adopted to enable the reworking of disciplinary knowledge generated during the study and analysis phase into the design action. The regeneration strategy developed was, in this sense, a pragmatization of the knowledge [47] and insights that emerged, as well as their reworking and organization into a coherent system capable of providing a framework and specific, detailed proposals for intervention.

Furthermore, the focus on community involvement and the design of the participatory process highlighted stakeholders' voices as a valuable asset. In the field of participatory planning, various solutions exist to prevent manipulation or the instrumental use of participants' goodwill. Some are adopted during the planning phase (establishing an advisory committee and denying promoters any form of power or privilege), and others during the implementation phase (providing participants with information, expertise, and balanced testimony, and entrusting process management to external parties). These solutions are commonly found in participatory projects, but they are not always completely successful [48].

The *Casa Sant'Angela* case study highlights that the use of an experimental method that offers tools and approaches to incorporate citizens' participation at different levels of the design process could prevent extractive and manipulative approaches toward citizens and participatory processes tout court.

This approach has made it possible to direct planning efforts toward representing—and thus legitimizing within the political process and institutions—the issues, concerns, and needs around which residents have rallied [49]. Rather than serving as a structure that conveys and informs, citizen participation has thus become an opportunity to discover the potential and highlight the needs of a community [50].

References

1. Bonfantini, G.B.: Progetto urbanistico e città esistente. Libreria Clup, Milano (2002)
2. Lefebvre, H., Kofman, E., Lebas, E.: Writings on Cities. Blackwell Publishers, Cambridge, Mass, USA (1996)
3. Caruso, N.: Atti della XXIII Conferenza Nazionale SIU - Società Italiana degli Urbanisti - DOWNSCALING, RIGHTSIZING. Contrazione demografica e riorganizzazione spaziale, Torino, 17–18 giugno 2021. Planum Publisher - Società Italiana degli Urbanisti, IT (2021)
4. Harvey, D.: The Urban Experience. Johns Hopkins Press, Baltimore (1989)

5. Peck, J.: Remaking laissez-faire. Prog. Hum. Geogr. **32**, 3–43 (2008). https://doi.org/10.1177/0309132507084816
6. Baeten, G.: Neoliberal planning. In: The Routledge Handbook of Planning Theory, pp. 105–117. Routledge, London (2017)
7. Culture Leading to Urban Regeneration. Empirical evidence from some italian funding programs. In: Lecture Notes in Networks and Systems, pp. 461–470. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-06825-6_43
8. Foglieni, F., Leoni, F., Cipriani, L., Maffei, S.: From ideas to policies, through places: service design-driven prototyping guidelines for urban regeneration. Presented at the ServDes.2023 Entanglements & Flows Conference: Service Encounters and Meanings Proceedings July 11 (2023)
9. Castet, N., Gouache, C.: Active citizens' action planning network. European Union, Bruxelles (2020)
10. Uitermark, J., Duyvendak, J.W., Kleinhans, R.: Gentrification as a governmental strategy: social control and social cohesion in Hoogvliet, Rotterdam. Environ Plan A. **39**, 125–141 (2007). https://doi.org/10.1068/a39142
11. Smith, N.: New globalism, new urbanism: gentrification as global urban strategy (2002)
12. Cocola-Gant, A., Lopez-Gay, A.: Transnational gentrification, tourism and the formation of 'foreign only' enclaves in Barcelona. Urban Stud. **57**, 3025–3043 (2020). https://doi.org/10.1177/0042098020916111
13. Lees, L.: Gentrification and social mixing: towards an inclusive urban renaissance? Urban Stud. **45**, 2449–2470 (2008). https://doi.org/10.1177/0042098008097099
14. Coppola, A., Granata, E., Lanzani, A., Longo, A.: C'è ancora spazio per l'urbanistica? Riflessioni a margine della vicenda "Salva Milano." Archivio di Studi Urbani e Regionali **55**, 5 (2025). https://doi.org/10.3280/asur2024-141001
15. Coppola, A., Pessina, G.: Le crisi territoriali tra mutamento dell'intervento statale, nuove strategie assicurative e nuove disuguaglianze Prospettive dagli Stati Uniti e dall'Italia: Prospettive dagli Stati Uniti e dall'Italia. In: Le disuguaglianze territoriali in Italia Cause, forme, conseguenze. p. Chapter_24346. Fondazione Università Ca' Foscari, Venice (2025). https://doi.org/10.30687/978-88-6969-991-7/014
16. Piletić, A.: The role of the urban scale in anchoring authoritarian neoliberalism: a look at post-2012 neoliberalization in Belgrade, Serbia. Globalizat. **19**, 285–300 (2022). https://doi.org/10.1080/14747731.2021.1882817
17. Dodds, R., Butler, R.: The phenomena of overtourism: a review. IJTC **5**, 519–528 (2019). https://doi.org/10.1108/IJTC-06-2019-0090
18. Arnstein, S.R.: A ladder of citizen participation. J. Am. Inst. Plann. **35**, 216–224 (1969). https://doi.org/10.1080/01944366908977225
19. Jacobs, J.: The death and life of great american cities. Random House (1961)
20. Cooper, D.G.: Participatory urban planning. In: Christens, B.D. (ed.) The Cambridge Handbook of Community Empowerment, pp. 241–261. Cambridge University Press (2024). https://doi.org/10.1017/9781009153720.012
21. Raynor, K., Doyon, A., Beer, T.: Collaborative planning, transitions management and design thinking: evaluating three participatory approaches to urban planning. Austral. Plan. **54**, 1 (2018). https://doi.org/10.1080/07293682.2018.1477812
22. Barry, J., Legacy, C.: Between virtue and profession: theorising the rise of professionalised public participation practitioners. Plan. Theory **22**, 85–105 (2023). https://doi.org/10.1177/14730952221107148
23. Reaven, M.: Citizen participation in city planning: New York City, pp. 1945–1975 (2009)
24. DiSalvo, C.: Design as Democratic Inquiry: Putting Experimental Civics into Practice. The MIT Press, Cambridge, Massachusetts (2022)

25. Marres, N., De Rijcke, S.: From indicators to indicating interdisciplinarity: a participatory mapping methodology for research communities in-the-making. Quant. Sci. Stud. **1**, 1041–1055 (2020). https://doi.org/10.1162/qss_a_00062
26. Bason, C.: Beyond labs: the future of civic design. Des. Manage. Rev. **33**, 32–39 (2022). https://doi.org/10.1111/drev.12296
27. Ehn, P.: Participation in design things. Presented at the Proceedings of the Tenth Anniversary Conference on Participatory Design 2008 (PDC 2008) (2008)
28. Dyer, M., Corsini, F., Certomà, C.: Making urban design a public participatory goal: toward evidence-based urbanism. Proc. Inst. Civil Eng. - Urban Des. Plan. **170**, 173–186 (2017). https://doi.org/10.1680/jurdp.16.00038
29. Karasti, H.: Infrastructuring in participatory design. In: Proceedings of the 13th Participatory Design Conference: Research Papers, pp. 141–150. Association for Computing Machinery, New York, NY, USA (2014)
30. Korn, M., Reißmann, W., Röhl, T., Sittler, D.: Infrastructuring Publics. Springer (2019)
31. Korn, M., Voida, A.: Creating friction: infrastructuring civic engagement in everyday life. AAHCC **1**, 12 (2015). https://doi.org/10.7146/aahcc.v1i1.21198
32. Pernecky, T. (ed.): Postdisciplinary Knowledge. Routledge, Taylor & Francis Group, London , New York, NY (2020)
33. Koskinen, I., Zimmerman, J., Binder, T., Redström, J., Wensveen, S.: Field. In: Design Research Through Practice, pp. 69–87. Elsevier (2012). https://doi.org/10.1016/B978-0-12-385502-2.00005-5
34. Findeli, A.: Design research-introduction. Des. Issues **15**(2), 1–3 (1999)
35. Findeli, A., Brouillet, D., Martin, S., Moineau, C., Tarrago, R.: Research through design and transdisciplinarity: a tentative contribution to the methodology of design research. Presented at the Swiss Design Network Symposium 2008 (2008)
36. Fareri, P.: Rallentare. Note sulla partecipazione dal punto di vista dell'analisi delle politiche pubbliche. In: Rallentare. Il disegno delle politiche urbane. Franco Angeli, Milano (2009)
37. Pasqui, G.: La città, i saperi, le pratiche. Donzelli, Roma (2018)
38. Manzini, E.: Making things happen: social innovation and design. Des. Issues **30**, 57–66 (2014)
39. Parente, M., Sedini, C.: Design for territories as practice and theoretical field of study. Des. J. **20**, S3047–S3058 (2017). https://doi.org/10.1080/14606925.2017.1352812
40. Tamborrini, P., Stabellini, B.: Metodologie e strumenti per l'innovazione sostenibile Il rilievo olistico come strumento per progettare sul territorio. MD J. (2018)
41. Manzini, E., Rizzo, F.: Small projects/large changes: participatory design as an open participated process. CoDesign **7**, 199–215 (2011). https://doi.org/10.1080/15710882.2011.630472
42. Bernstein, J.H.: Transdisciplinarity: a review of its origins, development, and current issues. J. Res. Pract. **11** (2015)
43. Piaget, J.: The epistemology of interdisciplinary relationship. In: Interdisciplinarity: Problems of Teaching and Research in Universities, pp. 127–139. OECD Publications Center, Whasington D.C. (1972)
44. Buchanan, R.: Wicked problems in design thinking. Des. Issues **8**, 5 (1992). https://doi.org/10.2307/1511637
45. Klein, J.T.: Interdisciplinarity and Complexity: An Evolving Relationship. Universidad Nacional Autónoma de México, México, D.F. (2004)
46. Rittel, H.W.J., Webber, M.M.: Dilemmas in a general theory of planning. Policy Sci. **4**, 155–169 (1973)
47. Cross, N.: Design research: a disciplined conversation. Des. Issues **15**, 5 (1999). https://doi.org/10.2307/1511837

48. Bobbio, L.: Designing effective public participation. Policy Soc. **38**, 41–57 (2019). https://doi.org/10.1080/14494035.2018.1511193
49. Latour, B.: What is the style of matters of concern? In: Gaskill, N., Nocek, A.J. (eds.) The Lure of Whitehead, pp. 92–126. University of Minnesota Press (2014). https://doi.org/10.5749/minnesota/9780816679959.003.0004
50. Latour, B., Weibel, P., Zentrum für Kunst und Medientechnologie Karlsruhe (eds.): Making things public: atmospheres of democracy ; [exhibition at ZKM, Center for Art and Media Karlsruhe 20.03. - 03.10.2005]. MIT Press [u.a.], Cambridge, Mass (2005)

APODigital: A Flexible Platform for Smart Surveys

Vinicio Bernardes Silva[1], Victor Guimarães[1], Laura Barros[1], Ana Julia Maria Silva[1], Diego Molinos[1], Simone Villa[1], and Flávio de Oliveira Silva[2(✉)]

[1] Federal University of Uberlândia (UFU), Uberlândia, Brazil
{vinicio.silva,victorguimaraes,laura.moreira1,anajuliamaria,diego.molinos,simonevilla}@ufu.br
[2] University of Minho, Braga, Portugal
flavio@di.uminho.pt

Abstract. Digital data collection is shifting from web-based questionnaires to smart surveys that integrate respondent feedback with multimodal device data; existing platforms often lack the flexibility required for complex, context-aware research. Existing platforms focus on specific use cases but do not support a broad range of scenarios. This work presents APODigital, a configurable smart-survey platform that extends traditional digital questionnaires by collecting additional data, such as camera and GPS location data, and by providing user-interacting questions generated by generative AI. The system design goal was to create a platform with a high degree of flexibility for creating questionnaires, while remaining simple enough for non-technical users to design and deploy customized survey applications. In addition to conventional question types, APODigital supports advanced features such as camera-based inputs, grouped evaluations, ranking-based feedback, image classification, and generative AI-assisted workflows. The system supports the design and deployment of smart surveys for flexible use across diverse scenarios. To showcase this flexibility, we present three applications built on the same platform: two housing-related evaluations and one health survey application. We demonstrate the platform's versatility through three distinct deployments: Meu Apê, a streamlined pilot involving 75 apartments; Como Você Mora?, a deep post-occupancy evaluation study with 246 participants; and a generic health questionnaire that showcases cross-domain reuse and the types of questions it supports. Our findings suggest that APODigital provides a scalable, reusable foundation for researchers requiring richer, context-driven data without the need for custom software development.

Keywords: smart surveys · multimodal data collection · post-occupancy evaluation · progressive web applications · generative AI

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 362–377, 2027.
https://doi.org/10.1007/978-3-032-30494-0_23

1 Introduction

Due to technological innovation, data collection has evolved from traditional web-based forms to *smart surveys*. In this type of survey, new devices, such as smartphones, wearables, and sensors, enable multimodal data acquisition, thereby enhancing the quality of depth data associated with users' responses [5]. In this type of survey, using sensors and artificial intelligence, respondents' answers are provided with improved context from additional data, thereby increasing awareness of the environment associated with these responses [14].

Despite the new possibilities that come with this type of survey, challenges remain, such as usability and a user experience that helps and engages the respondents, the use of Machine Learning (ML) and Natural Language Processing (NLP), and flexibility by using configurable, non-rigid systems that can integrate new technologies and handle diverse participant interaction methods [3]. Moreover, the use of generative AI in this area remains a challenge [15].

Post-Occupancy Evaluation (POE) aims to gather systematic feedback from building users about their experiences in the space and is inherently user-centered and data-driven [4]. As a tool capable of enhancing data gathering from the human interaction with the environment, *smart surveys* can help to enhance feedback from building users regarding their experiences within the space comfort, usability, sustainability, and environmental quality, thus helping to move from static data collection toward a dynamic, feedback-loop approach [4].

To address these limitations, we developed APODigital, a flexible platform that transforms standard questionnaires into multimodal *smart surveys* with adaptive paths and AI-driven interactions. Using a modular architecture, the system employs a Progressive Web Application (PWA) frontend managed through an administrative portal, enabling researchers to build customizable questionnaire flows that integrate with different generative AI models, allowing researchers without programming backgrounds to build and launch their own survey apps. The system handles common question types, such as text-based, single-choice, and multiple-choice, but also supports advanced features such as camera-based inputs, use of different device sensors, ranking systems, and generative AI-supported workflows that enrich the data collection process.

The contributions of this work are around the pillars. We present a configurable smart survey platform that leverages device sensors to enable multimodal data collection in questionnaires, supports camera- and GPS-based questions, provides different navigation flows based on conditions, and supports generative AI workflows. We detailed a layered, modular architecture that provides a responsive frontend across devices, a backend that supports structured and multimedia data handling, and integration with generative AI services using reusable backend components. Finally, we validated the system's versatility by deploying it across three real-world scenarios in different domains: two based on POE studies and one general survey that showcases the system's flexibility.

The remainder of this paper is organized as follows. Section 2 reviews related work on digital survey systems and smart surveys. Section 3 presents the architecture and design of the APODigital platform. Section 4 describes the imple-

mentation and the applications developed using the platform. Finally, Sect. 5 concludes the paper and outlines directions for future work.

2 Related Work

Recent research has increasingly shifted from traditional web-based questionnaires to *smart surveys*, i.e., survey systems that integrate self-reported responses with contextual data collected via smart devices. In this paradigm, smartphones are not merely used as form-filling interfaces but also as sensing platforms that capture contextual information through embedded hardware, such as cameras, GPS receivers, motion sensors, and other connected devices.

Schouten et al. [14] define smart surveys as survey systems that integrate respondent interactions with contextual information collected from digital devices. They state that the use of Artificial Intelligence and Machine Learning (AI/ML) remains challenging. [9] says that the use of sensors and machine intelligence can reduce respondent burden and increase data quality.

TigerAware, proposed by Morrison et al. [10], is a general-purpose, customizable mobile data-collection and analytics platform that supports both traditional survey responses and sensor-based data capture. TigerAware enables the collection of information from internal smartphone sensors, external sensing devices, and structured questionnaires, and provides real-time dashboards for visualization. As a tool presented in 2018, it does not use any AI-based technology.

Image-based responses in mobile surveys have also been explored by Bosch et al. [2]. Their study investigates the feasibility of allowing respondents to answer survey questions with photographs rather than text. They explore the use of images to obtain richer response context and a computer-vision-based API to support image-response classification at scale.

Conversational survey interaction has also been explored as an alternative survey interface paradigm. Building on *TigerAware*, Guo et al. [7] propose a serverless conversational survey system that enables researchers without programming skills to create voice-enabled chatbots for conducting customizable surveys. In this case, respondents interact with questionnaires via text or voice.

Giacalone et al. [5] present smart surveys that work together with IoT deployed in the environment where the survey takes place. They present a modular system that integrates IoT devices, environmental sensors, and wearable sensors. The system uses neural network-based components in the backend, but does not use generative AI to interact with the user. These deep interactions with IoT in environments hinder the system's widespread adoption across diverse areas.

To summarize APODigital's positioning in this landscape, Table 1 compares these works with respect to support for camera input, smartphone sensing, external sensor/IoT integration, adaptive questionnaire flow, and generative AI-supported interaction with respondents.

These studies presented in Table 1 show the ongoing transition from text-based digital questionnaires to smart-survey systems that incorporate data from

Table 1. Comparison of smart survey platforms and multimodal data collection capabilities

Work	Year	Cam	GPS	Ext	Adapt	Gen-AI	Main Contribution
Morrison et al. [10]	2018	Y	Y	Y	Y	N	TigerAware modular mobile survey and sensor data collection platform
Bosch et al. [2]	2019	Y	N	N	N	N	Image-based responses in mobile surveys with automated computer vision analysis
Guo et al. [7]	2021	N	N	N	Y	N	Conversational chatbot-based survey interaction
Alkassm et al. [8]	2022	Y	N	N	Y	N	Use of virtual reality environments to enhance questionnaire immersion
Rama et al. [13]	2022	N	N	N	Y	N	Innovative structuring and customization of Google Forms questionnaires
Bensfiya et al. [1]	2023	Y	N	N	L	N	Comparative study of digital assessment tools including Google Forms and QCM CAM
Giacalone et al. [5]	2025	Y	Y	Y	N	N	IoT-enabled smart survey architecture
APODigital	**2025**	**Y**	**Y**	**Y**	**Y**	**Y**	**Configurable multimodal survey platform integrating smartphone sensing capabilities and generative-AI supported interactions**

Columns: Cam = Camera input; GPS = Smartphone sensors (e.g., GPS); Ext = External sensors/IoT; Adapt = Adaptive questionnaire flow; Gen-AI = Generative AI-based questions.
Legend: Y = Supported; N = Not supported; L = Limited support; C = Conceptual framework

multiple media to enrich the context of responses, improve respondent engagement, and leverage advances in IoT and artificial intelligence.

While some tools focus on image [2] or voice [7] to enhance traditional surveys, or are tightly coupled with infrastructure [5], most systems lack a platform that integrates these capabilities into a unified survey infrastructure supporting surveys across different knowledge domains.

Although some tools have successfully integrated image [2] or voice [7] capabilities to enhance traditional surveys, and others have focused on heavy infrastructure [5], these solutions often lack a common thread. The current landscape lacks a truly integrated platform that brings these features together into a single, versatile system capable of supporting data collection across different knowledge domains.

To address this gap, APODigital is a configurable smart-survey platform that combines reusable questionnaire composition, multimodal interaction, adaptive navigation, and generative AI-based interactions with respondents. The platform's inception was motivated by POE. Still, we designed it as a general-purpose infrastructure capable of supporting a broader range of digital data-collection scenarios created by non-technical users.

3 APODigital Platform

This section introduces the Adaptive Platform for Omnimodal Digital Surveys (APODigital).

In addition to providing high flexibility for different types of surveys, another goal of the system design was to handle diverse data types, including user-entered text, sensor data, and images. To achieve this, we used a layered design pattern to support the system architecture. Figure 1 illustrates the architecture and overall workflow of the APODigital system.

The system is compatible with various client devices, and the user experience across all devices was also a key objective. The system can leverage the sensors available on these devices and request permission from the end user, in this case, the survey participant, to use them.

Initially, the system's end users design and configure questionnaires through the administrative portal. The system uses modern web development [11] and relies on a Vue-based frontend that consumes services from a RESTful backend API.

While the end-user creates new questions, the backend processes the requests and stores each question structure and its all related configuration in a PostgreSQL database.

Once a participant selects a questionnaire to answer, they launch a PWA. The frontend uses RESTful calls to load the required survey logic on the fly, which then dictates how the application renders prompts and navigation paths for that specific session.

By using a PWA, the system runs in standard web browsers, providing cross-platform compatibility without requiring a native application. This enables

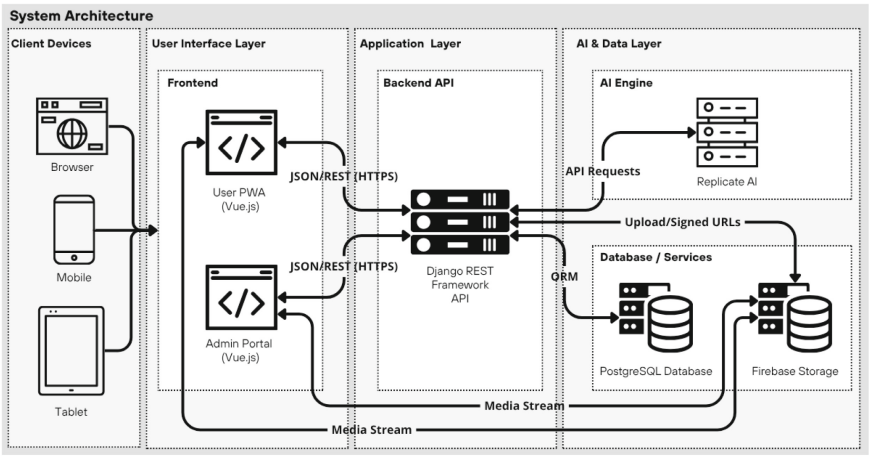


Fig. 1. APODigital Platform Architecture

participants to complete the questionnaires regardless of device type or operating system. As the participant provides responses, the data is transmitted to the backend, where it is validated and stored. Media data, such as images are uploaded directly to cloud storage services via secure mechanisms, such as signed URLs, thus reducing backend load.

The system integrates generative AI services via external APIs to support functionalities such as image generation and natural-language interaction. During participant responses, the system supports generative-AI-based interaction. This enhances the user experience and enables more advanced data-collection workflows.

The next subsections provide additional information about each system layer.

3.1 User Interface Layer

For the frontend, we chose the Vue framework, which enabled us to build a responsive interface that functions seamlessly as a PWA across desktops, smartphones, and tablets. Every questionnaire is assigned a unique URL, which the researcher defines within the administrative portal. When a respondent opens this link, the application dynamically retrieves the necessary configurations from the backend API to render the survey—a setup that provides the flexibility needed for diverse data-collection needs.

The *admin* portal serves as a no-code environment for researchers to design and manage their studies. Users can choose from a wide range of question types, including standard multiple-choice and open-ended fields, as well as more specialized tools such as grouped scales, camera-based inputs, and informational prompts. To make the surveys more interactive, we implemented client-side *go-to* logic that enables dynamic navigation, in which the question flow adapts

instantly based on a participant's previous answers. Beyond creation, the portal also supports cloning and editing, enabling existing studies to serve as templates for future campaigns and thereby significantly reducing setup time for new research.

We also integrated customizable startup screens to manage the first point of contact with participants. These screens allow researchers to clearly state the study's objectives, present privacy conditions, and obtain informed consent. This ensures that each respondent can review the participation agreement before opting in or declining.

Finally, the User PWA serves as the primary gateway for participants, designed for a smooth, intuitive experience. Recognizing that field research often happens in areas with poor service, we prioritized offline functionality. If a connection is lost, the PWA temporarily caches responses locally on the device. Once a stable network is detected, the system automatically synchronizes the data with the backend, ensuring no information is lost due to connectivity issues.

3.2 Application Layer

The Application Layer uses the Django framework, which provides the REST APIs that drive the main platform functions. It establishes the connection between the frontend, data persistence, and AI services, handling tasks such as questionnaire configuration, user session management, response validation, and the execution of core business rules. This layer is also responsible for managing database operations and integrating the platform with external services.

The services in this layer validate participant responses and securely store the data in a relational database. The backend also supports data retrieval and export. The researcher responsible for a questionnaire needs flexibility to work with the data. In this layer, the service logic exports results directly to formats such as Excel and Comma-Separated Values (CSV). This enables further statistical analysis in external tools, enhancing the system's flexibility.

This layer contains the core application logic. This approach improves future maintainability by providing a direct message flow between components and easing the addition of new functionality to the system.

3.3 AI & Data Layer – AI Engine

The AI & Data Layer is responsible for persistent storage, data processing, and the integration of AI-based services.

The *AI Engine* is a component that supports various AI capabilities, enabling the use of generative AI during questionnaire responses by providing generative AI-supported questions.

The generative-AI support currently exposed by the system is organized around two main functions. The first is image generation from textual prompts, which can be used during data collection or as part of the interaction flow presented to the user. The second is natural-language interaction with Large

Language Models (LLMs), enabling users to ask questions, seek clarification, and receive contextual assistance while using the platform.

To integrate these capabilities into the system, we used external APIs provided by the Replicate platform[1]. By doing so, we decouple the application logic from specific model implementations, simplifying the incorporation of new models as the system evolves. We also leave room for future evaluation of different AI models. The unified API sub-layer reduces the effort required to replace, update, or extend AI functionality while preserving the platform's overall organization.

Currently, image generation is supported by the Nano Banana model [6]. Natural-language assistance and image interpretation tasks are supported by the GPT-4o Mini model [12]. These services are invoked as part of the backend workflow and exposed to the application via controlled interfaces, enabling their incorporation without altering the system's higher-level structure.

In this work, the goal is to present the AI-enabled functions as integrated components of APODigital and illustrate their use through representative workflows in the generic application. Here, the main focus is on the system design and the integration of these capabilities into the platform architecture. As future work, we plan to conduct a systematic evaluation of these functionalities, including accuracy, usability, and their impact on the user experience across different AI models.

3.4 AI & Data Layer - Database and Services

The AI & Data Layer provides services for interacting with the database and storage services. The APODigital system needs to handle both structured and unstructured data.

The structured data includes questionnaire definitions, user responses, and sensor data. In our implementation, we use PostgreSQL. The media captured during the questionnaire responses constitute unstructured data that the system must handle. The design choice here was to store this type of data in a cloud-based service. By doing this, we decouple handling large files from the Application layer and improve scalability to support media-oriented workflows. In our implementation, images, geolocation, and generative-AI-signed URLs were used to access this data. This enables a more secure transfer between client devices and avoids routing large files to the backend.

The APODigital system also incorporates other mechanisms to enhance security and data governance for user interaction. These mechanisms include authenticated access, secure storage, and controlled transfer of both structured and unstructured data. This approach is important in application scenarios that simultaneously process images, geolocation, and generative AI.

4 Applications and Evaluation

To evaluate the practical applicability of APODigital, we examined its use in three questionnaire applications built on top of it. These are not three separate

[1] https://replicate.com/.

systems, but rather three different questionnaires that use a single smart-survey infrastructure. This evaluation enables us to assess how APODigital supports diverse questionnaire structures, interaction patterns, and data-collection needs.

From an evaluation standpoint, these applications highlight three complementary dimensions of the platform. They illustrate *domain-specific customization*, as shown by two POE questionnaires with different levels of complexity. They demonstrate support for *adaptive and multimodal interaction*, including conditional flows, camera-based inputs, grouped scales, ranking questions, and mobile-device features. Finally, they demonstrate cross-domain applicability and indicate that the infrastructure can support smart surveys across diverse scenarios.

This flexibility makes APODigital suitable for specialized applications in different research contexts. In this work, we use three applications to demonstrate this capability: *Meu Apê*, *Como Você Mora?*, and a *Generic Evaluation Application*. Together, they show that the platform can support both focused and more general-purpose data-collection scenarios, spanning different levels of questionnaire complexity and interaction richness.

4.1 Evaluation Perspective

The three applications were chosen to represent distinct usage profiles of the platform. *Meu Apê* focuses on streamlined interaction and quick completion. *Como Você Mora?* explores greater configurability and adaptive questionnaire behavior in a richer domain context. The *Generic Evaluation Application*, in turn, highlights cross-domain reuse and the incorporation of AI-enabled interaction features. For this reason, the evaluation presented in this section considers APODigital not as a single-purpose questionnaire tool, but as a reusable smart-survey infrastructure.

In the deployments involving real users, described in Subsect. 4.2 and Subsect. 4.3, participation was preceded by an explicit agreement screen. This ensured that users were informed about the study and could choose whether to participate before any data was collected.

Table 2 highlights how the same APODigital infrastructure supports different levels of questionnaire complexity, interaction richness, and domain adaptation without requiring changes to the underlying system architecture.

4.2 *Meu Apê*—Apartment-Focused Evaluation

The *Meu Apê* application, available at https://app-mora.facom.ufu.br/evaluation/meuape_tu, is a digital instrument for assessing the environmental quality of occupied apartment buildings. Its goal is to capture multiple dimensions of dwelling quality from physical, behavioral, and environmental perspectives, and to examine how residents' daily habits may help reduce environmental impacts.

From the perspective of platform evaluation, *Meu Apê* represents a *low-complexity deployment profile*. The questionnaire was designed to support quick

Table 2. Comparison of APODigital applications

Application	Domain	Flow	Multimodal	AI	Deployment	Evaluation focus
Meu Apê	Apartment POE	Linear	Medium	No	Pilot (75 apartments)	Low-complexity deployment; fast completion and high engagement
Como Você Mora?	Housing POE	Adaptive	Medium	No	Controlled study (246 questionnaires; 123 houses/123 apartments)	Complex branching, richer contextual data, and configurable survey logic
Generic Evaluation	Health	Mixed	High	Yes	Prototype	Cross-domain reuse and AI-enabled interaction

Columns: Flow = Questionnaire structure (Linear, Adaptive, Mixed);
Multimodal = Use of camera, grouped scales, ranking, and device-aware features;
AI = Use of AI-based features (image classification, generative workflows, LLM interaction);
Deployment = Observed usage context.

completion and encourage user engagement through a simplified structure and mostly linear navigation, with limited use of conditional logic. This configuration is particularly suitable for scenarios in which ease of use, short completion time, and higher response rates are important.

The survey is organized into eight sections: user profile, apartment characteristics, environmental comfort (thermal and lighting), energy efficiency, water efficiency, waste disposal habits, transport and carbon emissions, and new consumption habits. In addition to conventional questionnaire items, the application also leverages device-aware capabilities, such as the mobile camera and GPS, to enable more precise unit-level diagnostics.

The application was validated through an on-site pilot study involving residents of 75 apartments in the Imperial Residence condominium in Uberlândia, Brazil. This deployment provides initial evidence that APODigital can effectively support focused, real-world POE studies through a streamlined mobile data-collection process.

4.3 *Como Você Mora?*—Comprehensive Housing Evaluation

The *Como Você Mora?* application, available at https://app-mora.facom.ufu.br/evaluation/comovocemora, is a broader and more expressive tool designed to evaluate both houses and apartments. Compared with *Meu Apê*, it represents a *higher-complexity deployment profile*, requiring richer questionnaire logic and wider contextual coverage.

This application relies on more elaborate navigation structures and extensive conditional branching. As a result, the questionnaire can adapt dynamically to participants' responses and capture more detailed information, albeit at the cost of a longer completion time. From a platform perspective, this application is particularly relevant because it shows that APODigital can accommodate

adaptive questionnaire behavior and more sophisticated survey flows without requiring a different system implementation.

The survey is structured around six categories: "About You," "Previous Housing," "Current Housing," "Building and Complex," "Housing Unit," and "Sustainable Habits." In addition, the system includes nine informative screens that guide participants through the questionnaire, clarify technical issues, and provide feedback during the interaction.

The applicability of this configuration across different housing typologies was examined in a controlled study conducted in Uberlândia, Brazil, with 246 completed questionnaires, evenly divided between houses (123) and apartments (123). This study provides evidence that the same APODigital infrastructure can support more detailed evaluations, richer result structures, and distinct respondent profiles, further highlighting the platform's configurability and adaptability.

4.4 Generic Evaluation Application - Cross-Domain Reuse

Beyond housing evaluation, APODigital's modular design enables the creation of questionnaires across a wide range of research domains. To illustrate this flexibility, we developed a generic health questionnaire that demonstrates how the platform can be adapted to different research contexts while preserving the same core infrastructure. This questionnaire is available at https://app-mora.facom.ufu.br/evaluation/generic-english.

This application showcases several question types supported by the platform, including informational prompts, single-choice and multiple-choice questions, open-text responses, camera-based inputs, and grouped-scale evaluations. It also demonstrates more advanced AI-enabled features, such as image classification for automatic category identification and generative *img2img* models for interactive photo editing, both of which provide dynamic feedback based on user input.

The questionnaire includes the following interaction types:

- **Informational Question:** Introduces the questionnaire and explains the purpose of collecting data on health habits.
- **Single-Choice Question:** Allows respondents to select a single option, such as the frequency of physical activity.
- **Multiple-Choice Question:** Enables respondents to select multiple options, such as the types of food they regularly consume.
- **Text Question:** Provides an open field in which respondents can describe aspects of their physical health.
- **Camera-Based Question:** Allows respondents to submit a photograph related to an aspect of their well-being.
- **Grouped Scale Question:** Captures perceptions of different health aspects (e.g., sleep quality, diet, and exercise routine) using a scale ranging from "very good" to "very poor".
- **Generative Image:** Supports image editing through *img2img* models based on user-defined options.

- **Image Classification:** Uses artificial intelligence to identify categories from a user-provided image and provides immediate feedback based on the detected results.
- **Ranking Question:** Provides immediate *feedback* based on the responses, encouraging healthy behaviors and suggesting improvements when appropriate.

Figure 2a show examples of how a respondent sees the different types of question types that are commonly found in every questionnaire, namely informational, single-choice, multiple-choice, and text input. Figure 2b shows advanced types of questions, such as camera-based, grouped-scale, and ranking interactions, that enable real-time presentation of the score the respondent receives while answering some questions. For example, after several questions about eating habits, you can provide a score so that the respondent knows their rank in this regard.

Figure 3a presents the generative AI workflow based on *img2img* processing. This type of question uses the camera to enhance the generative AI-based interaction. Figure 3b shows the image-classification workflow with automated category identification and feedback provided to the respondent.

This application also shows that APODigital can be used beyond POE scenarios. The same infrastructure can be reused in a different domain while preserving its core architectural and interaction mechanisms, including dynamic rendering, question abstractions, multimodal support, and API-based AI integration.

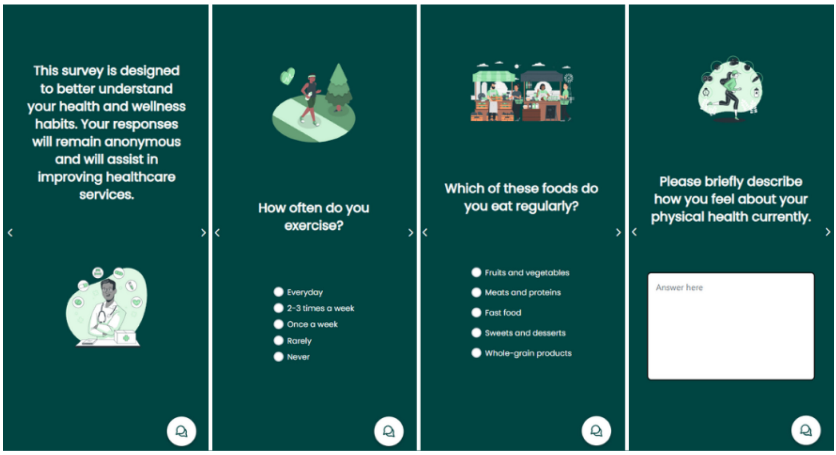
In this work, the goal is to present AI-enabled features supported by APODigital and to showcase the platform's extensibility and support for richer forms of interaction. A more detailed evaluation of their quality, reliability, and impact on users remains a topic for future work.

4.5 Discussion of Findings

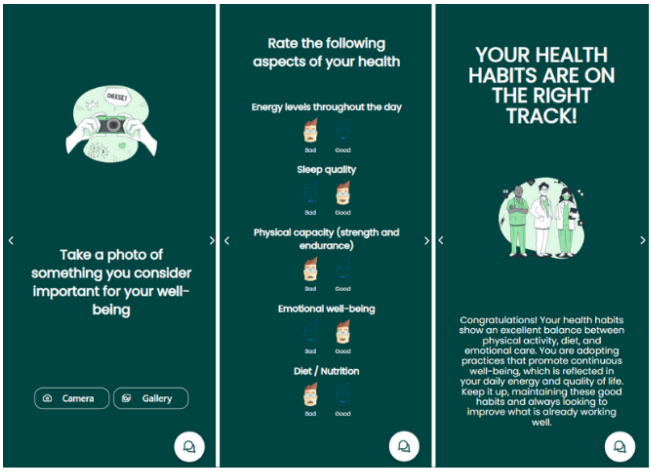
The three applications provide complementary evidence of APODigital's capabilities. First, the platform accommodates different levels of questionnaire complexity, ranging from mostly linear instruments designed for quick completion (*Meu Apê*, presented in Subsect. 4.2) to more elaborate surveys with extensive adaptive branching (*Como Você Mora?*, presented in Subsect. 4.3).

The system supports multimodal data collection through camera-based interaction and device-aware features, moving beyond conventional digital questionnaires toward richer smart-survey workflows.

The generic application exhibits that these capabilities can be reused in a very different context from the previous POE-based questionnaires, supporting APODigital's view of itself as a configurable smart-survey platform rather than a single-domain solution.



(a) Overview of basic question types: informational, single-choice, multiple-choice, and text input.

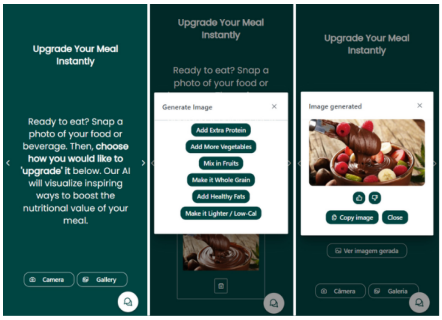


(b) Overview of advanced question types: camera-based, grouped scale, and ranking.

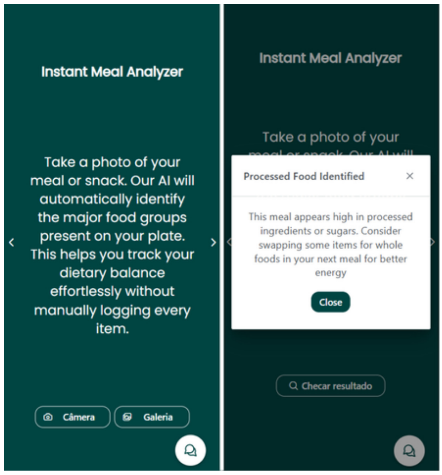
Fig. 2. Overview of the main application question types.

The *Admin Portal* supports configuring, adapting, and reusing questionnaires across applications. These features enhance the reuse of questionnaires with minimal technical intervention. We only described and briefly presented these features, while a full assessment was not the goal of this work.

The work has some limitations. The evidence presented here is based mainly on application deployments and functional demonstrations, rather than on a large-scale comparative user study.



(a) Generative AI workflow based on *img2img* processing.



(b) Image classification with automated category identification.

Fig. 3. Overview of the AI-assisted workflows.

In particular, this work does not compare APODigital with existing survey tools in terms of completion time, usability, or data quality. Similarly, although the generic application illustrates the integration of AI-enabled features, the present study does not provide a dedicated assessment of their effectiveness, reliability, or impact on user experience.

These aspects are important directions for future work and motivate a more comprehensive evaluation in subsequent studies.

5 Conclusion

This work presents APODigital, a configurable smart-survey platform that extends traditional digital questionnaires by incorporating multimodal data collection, adaptive navigation, and AI-enabled interactions. The platform combines a PWA frontend, a Django-based backend, and an AI & Data layer optimized to support both structured data from questionnaires and non-structured data from camera sensors.

Using an administrative portal, the APODigital's end user creates and deploys customized survey applications without requiring programming expertise. The questionnaire is a PWA-based application that can run on various device types and maintains a consistent user experience.

To showcase the platform's flexibility, we present three different questionnaires built on it. The two POE applications demonstrate how the platform can support both streamlined and more complex adaptive survey workflows. These POE applications were used in real-world assessments and remain online. While

currently primarily used in various POE scenarios, APODigital supports questionnaires in any area, and its generic evaluation, also available online, showcases its use across domains. The generic evaluation also demonstrates how APODigital's multimodal and AI-enabled capabilities can enable a new type of interaction that expands the smart-survey space.

These examples indicate that APODigital can serve as a reusable foundation for richer and more context-aware digital data collection. By integrating modern web technologies with advanced interaction types, such as camera-based input, grouped evaluations, ranking-based feedback, image classification, and generative AI workflows, the platform expands the design space of questionnaire-based systems beyond conventional web forms.

At the same time, the current study has limitations. The evaluation presented here is primarily based on application deployments and functional demonstrations rather than on a large-scale comparative user study.

In future work, we will focus on broader empirical validation, including usability analysis, completion-time assessment, and comparative studies with existing survey tools, as well as further exploration of privacy, ethics, and data quality issues in multimodal smart survey environments.

Acknowledgment. The authors gratefully acknowledge financial support from the Minas Gerais State Research Foundation (FAPEMIG), Brazil, under Grant #APQ-01727-24, and the institutional support provided by the Office of Research and Graduate Studies of the Federal University of Uberlândia (PROPP/UFU), Brazil.

References

1. Bensfiya, A., Boubih, S., Madrane, M., Janati-Idrissi, R.: Exploring digital alternatives for learning assessment: a comparative study of Google Forms and QCM CAM. Int. J. Sci. Didactics Educ. Eng. **1**, 44–55 (2023). https://doi.org/10.34874/PRSM.ijsdee-vol1iss1.877, https://journals.imist.ma/index.php/IJSDEE/article/view/877
2. Bosch, O.J., Revilla, M., Paura, E.: Answering mobile surveys with images: an exploration using a computer vision API. Soc. Sci. Comput. Rev. **37**(5), 669–683 (2019). https://doi.org/10.1177/0894439318791515
3. Bucher, H., et al.: Smart survey implementation. Workpackage 2: research methodology. Deliverable M6: review stage (2023)
4. Castaño-Rosa, R., Nerweyi, D.: Bringing post-occupancy evaluation up front to enhance energy efficiency in residential buildings. Build. Res. Inf. **53**(8), 1075–1091 (2025). https://doi.org/10.1080/09613218.2025.2538167
5. Giacalone, M., Marciano, C., Pipino, C., Piscopo, G., Marra, S.: Technological innovation and the role of smart surveys in the industrial context. Appl. Sci. **15**(16), 8832 (2025). https://doi.org/10.3390/app15168832, https://www.mdpi.com/2076-3417/15/16/8832
6. Google: Nano Banana image generation (2026). https://ai.google.dev/gemini-api/docs/image-generation

7. Guo, W., Zong, S., Chen, S., Zhao, F., Shang, Y.: Design and implementation of a new serverless conversational survey system. In: 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), October 2021, pp. 358–363 (2021). https://doi.org/10.1109/ICDSCA53499.2021.9650203, https://ieeexplore.ieee.org/abstract/document/9650203
8. Kassm, M.A.: The virtual reality questionnaire watch: exploring novel methods integrated with Google Forms. Thesis, Carleton University, 19 September 2022. Supervisor: Rob J. Teather. https://repository.library.carleton.ca/files/p8418p298
9. Mccool, D., Lugtig, P., Struminskaya, B.: Integrating smart surveys with traditional surveys. arXiv arXiv:2510.07521 [stat], October 2025. https://doi.org/10.48550/arXiv.2510.07521, http://arxiv.org/abs/2510.07521
10. Morrison, W., Guerdan, L., Kanugo, J., Trull, T., Shang, Y.: TigerAware: an innovative mobile survey and sensor data collection and analytics system. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), June 2018, pp. 115–122 (2018). https://doi.org/10.1109/DSC.2018.00025, https://ieeexplore.ieee.org/abstract/document/8411846/authors
11. Ollila, R., Mäkitalo, N., Mikkonen, T.: Modern web frameworks: a comparison of rendering performance. J. Web Eng. **21**(3), 789–813 (2022). https://doi.org/10.13052/jwe1540-9589.21311, https://ieeexplore.ieee.org/abstract/document/10243623
12. OpenAI: GPT-4o mini Model API (2026). https://developers.openai.com/api/docs/models/gpt-4o-mini
13. Rama, K., Sudha, R., Prasad, G.: An innovative way of asking questions using google forms. NeuroQuantology **20**, 20–1477 (2022). https://doi.org/10.48047/NQ.2022.20.20.NQ109150
14. Schouten, B., Lugtig, P., Luiten, A.: Can smart surveys have a positive business case? An evaluation based on three case studies. J. Off. Stat. **41**(2), 547–568 (2025). https://doi.org/10.1177/0282423X251321634
15. Valenzuela, S., Winter, S., Rivera, S.: Using large language models for survey research in communication: opportunities and challenges. Commun. Change **1**(1), 14 (2025). https://doi.org/10.1007/s44382-025-00014-z

A Thompson Sampling Based Operator Selection in Adaptive Large Neighborhood Search for Facility Location with Customer Incompatibilities

Gabriel Souto[1(✉)], Diego Brandão[2], Luidi Simonetti[1], and Pedro Henrique González Silva[1]

[1] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
{soutog,luidi,pegonzalez}@cos.ufrj.br
[2] Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil
diego.brandao@cefet-rj.br

Abstract. The Multi-Source Capacitated Facility Location Problem with Customer Incompatibilities (MS-CFLP-CI) arises in real-world decision-making contexts such as urban logistics, healthcare service planning, and hazardous material distribution, where operational constraints prevent certain customers from being served by the same facility. These incompatibility constraints significantly increase problem complexity and limit the applicability of exact optimization methods, especially in large-scale scenarios. In this work, we address this challenging problem by proposing an Adaptive Large Neighborhood Search (ALNS) framework enhanced with a learning-based operator selection mechanism. Specifically, we integrate contextual Thompson Sampling to dynamically balance exploration and exploitation based on the current search state. In addition, we introduce a novel conflict-aware destroy operator that explicitly targets regions of the solution space characterized by high incompatibility interactions. The proposed approach is designed to support large-scale decision-making, remaining scalable to instances involving thousands of facilities and customers. Computational experiments on benchmark datasets demonstrate that the method consistently improves both solution quality and robustness compared to state-of-the-art heuristics under comparable time limits. These results indicate that combining adaptive search with learning-based operator selection provides an effective and scalable strategy for solving complex facility location problems with real-world constraints.

Keywords: Large Neighborhood Search · Bandit Algorithms · Adaptive Operator Selection · Facility Location · Combinatorial Optimization

1 Introduction

Facility location problems play a central role in a wide range of real-world decision-making systems, including urban logistics, healthcare service planning,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 378–395, 2027.
https://doi.org/10.1007/978-3-032-30494-0_24

waste management, and hazardous material distribution [10]. In these applications, planners must determine where to open facilities and how to allocate demand while respecting operational constraints, regulatory requirements, and cost efficiency. These decisions directly impact service quality, safety, and resource utilization, making facility location a key component of modern decision-support systems [3,7].

A challenging aspect in many of these settings is the presence of incompatibility constraints between customers. Such constraints arise when certain entities cannot be jointly served by the same facility due to safety regulations, risk mitigation policies, or operational conflicts. For example, hazardous and non-hazardous materials must often be separated, and competing clients may require exclusive service conditions. These incompatibilities introduce a highly combinatorial structure into the problem, significantly increasing its complexity and limiting the effectiveness of traditional exact optimization methods, especially in large-scale scenarios.

The Multi-Source Capacitated Facility Location Problem with Customer Incompatibilities (MS-CFLP-CI) extends MS-CFLP by prohibiting specific customer pairs from being served by the same facility [7]. This additional disjunctive structure makes even the allocation subproblem NP-hard and leads to large, challenging benchmark instances [3,4]. Recent studies report strong metaheuristics and neighborhood-based methods, but solution quality remains sensitive to operator choices and instance characteristics [3,4,7].

These characteristics make the MS-CFLP-CI relevant for decision-support systems, where planners must evaluate complex trade-offs between cost, capacity, and operational feasibility. In such contexts, scalable and adaptive optimization methods are essential to support high-quality decisions in large-scale and constraint-rich environments.

In this work, we address this challenging optimization problem by proposing an Adaptive Large Neighborhood Search (ALNS) framework for MS-CFLP-CI that combines a new conflict-aware destroy operator (CCF) with a learning-based operator selection mechanism based on Thompson Sampling [11,13]. The goal is to balance exploration and exploitation while adapting the search behavior to the evolving solution structure and instance characteristics, ultimately supporting more effective decision-making in large-scale facility location problems.

The remainder of this paper is organized as follows: Sect. 2 defines the problem and its formulation. Section 3 reviews related work on MS-CFLP-CI. Section 4 describes the proposed ALNS and the CCF operator with Thompson Sampling. Section 5 reports the computational experiments. Finally, Sect. 6 concludes the paper.

2 Problem Definition

For a mathematical definition of MS-CFLP-CI, we consider a set of customers $N = \{1, \ldots, n\}$ and a set of candidate facilities $M = \{1, \ldots, m\}$. Each customer

$i \in N$ has demand $d_i > 0$, each facility $j \in M$ has capacity $s_j > 0$ and fixed opening cost $f_j > 0$, and c_{ij} denotes the unit shipping cost from facility j to customer i. Let $\mathcal{I}$ be the set of incompatible customer pairs; for each $(i_1, i_2) \in \mathcal{I}$, customers i_1 and i_2 cannot be served by the same facility. In the multi-source setting, a customer may be supplied by multiple facilities.

We use continuous variables $x_{ij} \in [0, 1]$ to represent the fraction of customer i's demand supplied by facility j (the shipped quantity is $d_i x_{ij}$). Binary variables y_j indicate whether facility j is open, and binary variables w_{ij} indicate whether facility j supplies customer i with a positive amount. The MS-CFLP-CI can be modeled as the following Mixed Integer Programming [3,4,7]:

$$\min \sum_{i \in N} \sum_{j \in M} c_{ij} d_i x_{ij} + \sum_{j \in M} f_j y_j \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in M} x_{ij} = 1 \qquad \forall i \in N \tag{2}$$

$$\sum_{i \in N} d_i x_{ij} \leq s_j y_j \qquad \forall j \in M \tag{3}$$

$$x_{ij} \leq w_{ij} \qquad \forall i \in N, \forall j \in M \tag{4}$$

$$w_{ij} \leq y_j \qquad \forall i \in N, \forall j \in M \tag{5}$$

$$w_{i_1 j} + w_{i_2 j} \leq 1 \qquad \forall (i_1, i_2) \in \mathcal{I}, \forall j \in M \tag{6}$$

$$x_{ij} \in [0, 1],\ w_{ij} \in \{0, 1\},\ y_j \in \{0, 1\} \qquad \forall i \in N, \forall j \in M \tag{7}$$

The objective (1) minimizes the total opening and transportation costs. Constraints (2) ensure that each customer's demand is satisfied, while Constraint (3) enforces facility capacities and implicitly prevents closed facilities from shipping. Constraints (4) and (5) link the continuous allocation variables to the binary variables, and (6) enforces customer incompatibilities for each facility. The linking Constraint (5) is redundant given Constraint (3), but it is commonly included to strengthen the formulation and improve solver performance [4].

3 Related Works

The MS-CFLP-CI was introduced by [7] along with the first public benchmark instances (wlp). The authors called the problem as a multi-source capacitated location model with incompatibility pairs and proposed a portfolio of independently developed metaheuristics. The portfolio includes MineReduce, a multistart ILS enriched with data mining, a GRASP method, a permutation-coded evolutionary algorithm, and a multistart greedy approach.

Their experiments used a common hardware platform and time limits tied to instance size to ensure fair comparisons. Results indicate that MineReduce is generally the best method across most instances, while the simpler greedy method can be competitive on the very large cases. Beyond algorithmic performance, the work is notable for providing instance generators, a solution validator, and open implementations, enabling direct reproducibility and comparisons.

Ceschia and Schaerf [3] proposed a multi-neighborhood simulated annealing (SA) approach that targets large instances through search space reductions. Their method starts from a greedy initial solution and explores neighborhoods that change suppliers, swap suppliers between customers, and close/open facilities. To control complexity, they restrict solutions to at most two suppliers per customer and limit candidate suppliers to a short list, which reduces the cost of feasibility checks.

The SA method is tuned with F-Race and tested on the wlp dataset and on a new benchmark (cflp-ci) released by the authors. The solver improves upon prior methods, achieving average improvements above 6% relative to MineReduce on most instances. The paper also provides models and lower bounds for smaller instances, strengthening the empirical baseline for MS-CFLP-CI.

Gjergji et al. [4] propose the first Large Neighborhood Search (LNS) designed for the MS-CFLP-CI. Their method introduces a set of problem-aware destroy operators that explicitly account for customer incompatibilities, coupled with a repair phase formulated as a restricted mixed-integer program. By embedding an exact solver within the heuristic loop, the approach effectively combines the intensification of mathematical programming with the flexibility of large-scale neighborhood exploration.

Computational experiments on benchmark datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches in terms of both solution quality and robustness under comparable time limits.

The observed variability in operator effectiveness motivates the use of adaptive operator selection mechanisms. In addition, we propose a new destroy operator, CCF, designed to exploit incompatible customers. This observation aligns with the objective of the present work, which extends the ALNS framework by integrating Thompson Sampling to dynamically balance exploration and exploitation, allowing the search process to adapt online to instance-specific characteristics.

These works highlight the importance of developing scalable optimization methods to address real-world facility location problems with complex operational constraints.

4 Proposed Method

This section presents the proposed methodology, which extends the Large Neighborhood Search (LNS) framework for the MS-CFLP-CI in two directions. First, we introduce a new conflict-aware destroy operator (CCF) that explicitly exploits the incompatibility structure of the problem. Second, we incorporate a learning-based operator selection mechanism based on contextual Thompson Sampling to dynamically guide the search process.

We begin by describing the set of destroy and repair operators (Sect. 4.1), followed by the Adaptive Large Neighborhood Search (ALNS) framework (Sect. 4.2), and finally the contextual Thompson Sampling mechanism used for operator selection (Sect. 4.3).

4.1 Destroy and Repair Operators

The proposed method relies on a set of destroy and repair operators that define large neighborhoods explored during the search. We adopt the operators introduced by [4] as a baseline and extend this set with a new conflict-aware operator.

Operators from Gjergji et al. We consider the three destroy operators proposed in [4], namely CW (Cheapest Facilities), CC (Cheapest Customers), and EC (Expensive Customers), together with an MIP-based repair operator.

The CW operator computes, for each open facility f', the average cost of serving the customers currently assigned to another facility f, and selects the ℓ facilities with the lowest values. The customers assigned to these facilities define the subproblem.

The CC operator selects the ℓ customers with the lowest unit shipping costs and includes these customers along with their associated facilities in the subproblem.

The EC operator builds a pool consisting of the most expensive customer of each open facility, selects one of them (either randomly or based on cost), and then includes the ℓ facilities with the lowest cost for that customer, along with their associated customers.

Given the subset of customers selected by a destroy operator, the repair phase consists of solving a restricted mixed-integer programming (MIP) model that preserves all original constraints while allowing customer reassignment and facility open/close decisions within the selected subset.

Conflict-Aware Customer Facility (CCF) Operator. While the operators of [4] are driven by cost-based criteria, they do not explicitly exploit the incompatibility structure of the MS-CFLP-CI. However, infeasibility and poor solution quality are often driven by strongly coupled, incompatible customers.

To address this limitation, we propose a conflict-aware destroy operator, named **CCF**, that targets regions of the solution space with high incompatibility density.

The operator assigns a score to each customer i defined as:

$$\text{score}(i) = \alpha \cdot \text{costShare}(i) + (1 - \alpha) \cdot \deg_I(i), \tag{8}$$

where $\text{costShare}(i)$ represents the fraction of total transportation cost associated with customer i, and $\deg_I(i)$ denotes the normalized degree of i in the incompatibility graph.

A seed customer is selected with probability proportional to this score. The subproblem is then expanded as:

$$U = \{i\} \cup N_I(i) \cup \text{coServed}(i), \tag{9}$$

where $N_I(i)$ represents the set of incompatible neighbors of i, and $\text{coServed}(i)$ denotes customers currently sharing facilities with i.

If the size of U exceeds a predefined limit ν, the set is truncated by prioritizing high-degree incompatible neighbors and selecting additional customers based on their score.

This construction allows the repair phase to focus on highly constrained regions of the solution, enabling more effective restructuring of assignments under incompatibility constraints.

4.2 Adaptive Large Neighborhood Search Framework

Given the set of destroy and repair operators described above, we embed them into an Adaptive Large Neighborhood Search (ALNS) framework [11].

At each iteration, the algorithm selects a destroy operator to define a subproblem, which is then reconstructed by the repair operator. This process allows the exploration of large and adaptive neighborhoods, balancing intensification and diversification.

Starting from an initial solution s_0, generated using the constructive heuristic of [4], the algorithm iteratively applies destroy and repair steps to produce candidate solutions.

Acceptance of candidate solutions follows a simulated annealing criterion, allowing occasional worsening moves with probability:

$$\exp\left(-\frac{f(s') - f(s)}{T}\right),$$

where $f(\cdot)$ denotes the objective value and T is the current temperature.

In contrast to classical ALNS, where operator selection is governed by segment-based weight adaptation and roulette-wheel selection, we replace this mechanism with a learning-based approach described in the next section.

4.3 Contextual Thompson Sampling for Operator Selection

The effectiveness of destroy operators may vary significantly across search phases and instance characteristics. The adaptive weight mechanism described in the ALNS framework of [11], in which operators are estimated from aggregated rewards over fixed-length segments, selects operators using a roulette wheel. While effective and widely adopted, adaptation occurs only at segment boundaries, and operator choice is independent of the current search state. As a result, the method can react slowly to changes in search dynamics and does not explicitly exploit context.

To address this, we replace the segment-based roulette selection with a learning-based mechanism using contextual Thompson Sampling [1,2,6,8,12,13]. In our implementation, each arm corresponds to a *pair* (destroy, repair); however, for the TS runs reported here, we fix the repair to MIP-only and use the (CW, CC, EC, CCF) destroy operators, so each arm corresponds to one destroy operator. We also model operator selection as a linear contextual bandit problem, in which the expected reward of each arm is assumed to be a linear function of the current search context.

Let $x \in \mathbb{R}^d$ denote the context vector, where d is the number of contextual features used to summarize the current search state.

$$x = [1,\ \text{progress},\ T/T_0,\ \text{no_improve},\ \text{last_acc},\ \text{last_imp},\ \text{infeas},\ \text{open_ratio},\ \text{conflict_ratio}] \tag{10}$$

where progress is elapsed time over the time limit, T/T_0 is the normalized temperature, no_improve is the normalized iterations since last best, last_acc and last_imp capture the previous acceptance and relative improvement, infeas flags the last infeasible attempt, open_ratio is the fraction of open facilities, and conflict_ratio is a fraction of facility–customer pairs that are blocked by incompatibilities in the current allocation solution. For each arm a, the expected reward is modeled as

$$\mathbb{E}[r \mid x, a] = x^\top \theta_a,$$

with an independent Bayesian linear model per arm. At each ALNS iteration, we sample $\tilde{\theta}_a$ from the posterior of each arm and select the arm maximizing $x^\top \tilde{\theta}_a$.

The selected features capture aspects of the search dynamics, including progress in time, acceptance behavior, solution structure, and conflict intensity, enabling TS to condition operator choice on search phase and problem-specific difficulty.

The reward used in our TS runs is binary: if the candidate is accepted and yields either a new global best or an improvement over the current solution of at least a relative threshold τ, the reward is 1; otherwise, it is 0. If a candidate is rejected or the destroy/repair fails, the reward is 0. To cope with non-stationarity, posterior updates apply exponential forgetting with factor ρ. We adopt a binary reward for simplicity and robustness, so that operator credit reflects whether an iteration produced a meaningful accepted improvement.

It is important to emphasize that the proposed Thompson Sampling mechanism replaces only the operator selection and adaptation components of the baseline ALNS. All remaining components of the baseline algorithm, namely the MIP-based repair procedure, the simulated annealing acceptance criterion, and the overall iterative search loop, are left unchanged. This design choice ensures that the proposed method preserves the structure of ALNS while introducing a context-aware operator selection mechanism capable of adapting online to the evolving search dynamics.

Algorithm 1 presents the resulting ALNS-TS algorithm as implemented in our experiments. The initial solution s_0 is generated using the same constructive heuristic employed by the LNS, which opens facilities according to fixed costs and assigns demands greedily. Within the Thompson Sampling framework, each arm corresponds to a pair in the Cartesian product $D \times R$ of destroy and repair operators. Since the repair set R contains only the MIP-based repair in our implementation, each arm effectively represents a distinct destroy operator (CW, CC, EC, and CCF). At each iteration, operator selection is conditioned on a contextual feature vector x computed from the current search state, and the algorithm proceeds until the time limit t is reached.

Algorithm 1: ALNS-TS

Input: Initial solution s_0, destroy set D, repair set R (MIP), time limit t
Output: Best solution s^*
Build s_0 with the LNS constructive heuristic; initialize posterior parameters for each arm $a \in D \times R$; $s \leftarrow s_0$, $s^* \leftarrow s_0$;
while *time* $< t$ **do**
  Compute context vector x;
  Sample $\tilde{\theta}_a$ for each arm $a = (d, r) \in D \times R$ and select $a = \arg\max_a x^\top \tilde{\theta}_a$;
  Apply destroy operator d to s and build subproblem $(U, F^{open}, F^{closed})$;
  Apply repair operator r to obtain candidate s';
  Compute reward q from the outcome of s';
  Update posterior of arm a with reward q and forgetting factor ρ;
  accepted $\leftarrow$ simulated-annealing decision on s' given s;
  if *accepted* **then**
    if $f(s') < f(s^*)$ **then**
      $s^* \leftarrow s'$;
    $s \leftarrow s'$;
return s^*;

In this way, operator selection becomes adaptive at every iteration, allowing the search to react online to changes in solution quality, stagnation, and incompatibility structure.

5 Computational Experiments

This section summarizes the experimental setup and evaluation methodology, then reports results in three parts: experimental details and benchmark instances, an ablation analysis to isolate the contribution of Thompson Sampling, and a comparison with the state-of-the-art LNS, followed by a statistical significance analysis of the observed improvements.

5.1 Experimental Details and Instances

We follow the evaluation methodology of [4] and use the two public benchmark instance sets for MS-CFLP-CI. The `wlp` dataset from [7] contains 30 instances, ranging from small cases (50 facilities) to very large ones (up to 3000 facilities). The `cflp-ci` dataset from [3] includes 50 instances with sizes up to roughly 3000 facilities, providing a complementary and challenging test bed. Each instance is solved under two timeouts: a short timeout of $10\sqrt{m}$ seconds and a long timeout of m seconds, where m is the number of facilities.

For ALNS-TS, we use the destroy operators CW, CC, EC, and the conflict-aware CCF, with MIP-only repair. Parameter values were selected via a grid-search procedure over the main ALNS-TS hyperparameters. The repair subproblem is solved with a time limit of 20 s per call, a maximum of two extra facilities

opened $(o + 2)$, and a relative MIP gap of 10^{-4}, using a single CPLEX thread. The subproblem size follows the dynamic rule $\ell = \nu|F|/n$ with $\nu = 50$ (Eq. (12) in 4), and closed facilities are selected using the same probability $p_{\text{MC}} = 0.5$ as in the LNS operators. Simulated annealing acceptance is enabled, with a time-based schedule and final temperature ratio of 0.05; the initial temperature is set to accept a 1% worsening with probability 0.5.

For contextual Thompson Sampling, we use a binary reward with success threshold $\tau = 0.001$, a Gaussian prior with $\lambda = 1.0$, noise $\sigma = 0.2$, and forgetting factor $\rho = 0.995$, with the `basic` context feature set. All methods use the same constructive heuristic (open facilities by fixed cost and assign demands greedily) and are executed on a single CPU core for a fair comparison. All experiments were run on a single machine with 16 GB RAM and an Intel Core i7-11800H (8 cores/16 threads, 2.30 GHz), x86_64 architecture.

The computational overhead introduced by the contextual Thompson Sampling mechanism is negligible compared to the cost of the MIP repair phase. Per iteration, TS requires only the construction of a low-dimensional context vector $(d = 9)$, one posterior sample per arm, and an inner-product evaluation, which together take well under one millisecond. Since each MIP repair call is bounded by 20 s and typically dominates the iteration time, the relative overhead of the learning mechanism remains below 0.1% of total runtime in our experiments.

5.2 Ablation Analysis: Impact of Thompson Sampling

In order to investigate the contribution of the proposed Thompson Sampling–based operator selection mechanism, we conduct an ablation study. We compare two algorithmic variants: the Adaptive Large Neighborhood Search without learning-based control (ALNS), and the proposed method ALNS-TS, which replaces the classical weight adaptation and roulette-wheel selection by contextual Thompson Sampling.

The evaluation is conducted on a subset of 15 instances randomly selected from the full benchmark set of 80 instances. Each instance is solved independently five times using a long time limit of m seconds, where m denotes the number of facilities in the instance. For each method, we report both the best solution obtained across runs (*best*) and the average solution quality (*avg*), allowing us to assess not only peak performance but also robustness.

Table 1 reports the best solution values obtained by ALNS, and ALNS-TS across the 15 tested instances. Overall, ALNS-TS attains the best solution quality on the majority of instances, outperforming the baseline ALNS.

In particular, ALNS-TS is the method capable of reaching the best known solution for instance `wlp01`, highlighting its enhanced ability to identify high-quality regions of the search space. For instance `wlp21`, ALNS-TS and ALNS tie by both reaching the optimal solution, indicating that the learning-based operator selection does not degrade performance when the baseline ALNS is already effective.

On a small number of instances (e.g., `cflp_ci_19`, `cflp_ci_48`, and `wlp15`), ALNS achieves marginally better best solutions than ALNS-TS. However, the

Table 1. Best and average solution values: ALNS and ALNS-TS.

Instance	Best		Average	
	ALNS	ALNS-TS	ALNS	ALNS-TS
cflp_ci_06	152.710	**152.650**	152.894	**152.828**
cflp_ci_10	600.146	**600.060**	600.458	**600.391**
cflp_ci_11	**30.739**	**30.739**	30.758	**30.783**
cflp_ci_19	**149.900**	149.993	**150.096**	150.097
cflp_ci_28	463.140	**463.111**	463.411	**463.302**
cflp_ci_39	189.904	**189.892**	189.982	**189.976**
cflp_ci_48	**765.460**	765.774	**765.604**	766.018
wlp01	28.730	**28.716**	28.742	**28.728**
wlp02	53.120	**52.975**	53.181	**53.146**
wlp03	64.415	**64.395**	64.508	**64.607**
wlp12	297.542	**297.349**	297.661	**297.535**
wlp15	**457.249**	457.251	**457.477**	457.535
wlp17	562.406	**562.386**	562.656	**562.633**
wlp21	**38.067**	**38.067**	38.150	**38.112**
wlp28	455.931	**455.826**	456.201	**456.032**

absolute differences are small when compared to the scale of the objective values, remaining within a small range even for large-cost instances. This indicates that the integration of Thompson Sampling does not introduce bias and remains competitive even in unfavorable cases.

The average solution values reported in Table 1 reveal an advantage of ALNS-TS over the ALNS. Across the ablation set, ALNS-TS improves the average solution in 12 out of 15 instances and never degrades performance by more than 0.05%.

Notably, ALNS-TS consistently improves upon ALNS, even when differences in best solution quality are marginal. This behavior suggests that the learning-based operator selection mechanism leads to a more stable search process, mitigating performance variability induced by random decisions and initial conditions. The gains in average performance are particularly relevant under the considered long time limit of m seconds, as they reflect sustained efficiency rather than isolated successful runs.

On instances where ALNS attains slightly better average values (e.g., `cflp_ci_19`, `cflp_ci_48`, and `wlp15`), the observed gaps remain small in absolute terms and do not outweigh the overall robustness gains achieved by ALNS-TS across the test set. Overall, the results indicate that contextual Thompson Sampling enables a better balance between intensification and diversification by adapting operator selection at every iteration based on the current search context, rather than relying on delayed, segment-based updates.

Overall, the ablation study demonstrates that the performance gains observed in ALNS-TS stem directly from the integration of Thompson Sampling into the operator selection mechanism. The method consistently improves average performance and frequently improves best-known solutions, without introducing instability or significant performance regressions. These results confirm that learning-based, context-aware operator selection constitutes an enhancement over ALNS schemes in MS-CFLP-CI.

5.3 Comparison with State-of-the-Art Methods

Based upon the ablation analysis, we now evaluate the ALNS-TS method against the current state of the art for MS-CFLP-CI, the LNS of [4]. Under the same two timeouts ($10\sqrt{m}$ and m seconds). Tables 2 and 3 report the best (min) and average (avg) solution values over the ten runs for each instance. The results show a clear and consistent advantage for ALNS-TS in both datasets, in terms of both best and average costs.

For `wlp` (Table 2), ALNS-TS improves the best solution in 26/30 instances under the short timeout and 28/30 under the long timeout; for average solutions, ALNS-TS is better in 29/30 instances under both budgets. On average, the relative improvement of ALNS-TS over LNS is about 0.79% (best, short), 0.88% (avg, short), 0.99% (best, long), and 1.09% (avg, long).

For `cflp-ci` (Table 3), the dominance is even stronger: ALNS-TS improves the best solution in 49/50 instances under the short timeout and 50/50 under the long timeout, and improves the average solution in 49/50 instances under the short timeout and 50/50 under the long timeout. The mean relative gains are approximately 0.76% (best, short), 0.94% (avg, short), 0.20% (best, long), and 0.24% (avg, long).

To compare methods on a common scale, we compute the percentage gap to the best-known solutions (BKS) as in [4]:

$$\text{GAP}(\%) = \frac{Z_{\text{sol}} - \text{BKS}}{\text{BKS}} \cdot 100, \tag{11}$$

where Z_{sol} is the solution value (min or avg) for a given method and timeout. We compare LNS and ALNS-TS using the same BKS values per instance, as shown in Table 2 and 3.

Figure 1 shows the GAP values for all 80 instances (30 `wlp` + 50 `cflp-ci`) under the m seconds timeout, demonstrating both best and average solutions for LNS and ALNS-TS. The figure highlights two patterns: (i) ALNS-TS shifts the GAP curves downward for both min and avg, indicating improved solution quality relative to LNS, and (ii) the separation between the average curves is stable across instances, evidencing reduced variance and improved robustness.

These improvements suggest that the proposed approach is suitable for supporting large-scale decision-making in real-world facility location systems with complex operational constraints.

Table 2. Results for dataset `wlp`, both timeouts.

Inst	$10\sqrt{m}$ seconds timeout				m seconds timeout			
	LNS		ALNS-TS		LNS		ALNS-TS	
	min	avg	min	avg	min	avg	min	avg
wlp01	28768.00	28805.30	**28716.00**	**28720.67**	28768.00	28805.30	**28716.00**	**28716.00**
wlp02	53133.50	53247.90	**52987.00**	**53011.33**	53133.50	53247.90	**52990.00**	**53017.75**
wlp03	**64482.50**	**64641.80**	64594.00	64754.25	**64482.50**	**64641.80**	64504.00	64751.00
wlp04	85039.50	85159.10	**84831.00**	**84881.42**	85039.50	85159.10	**84714.00**	**84723.58**
wlp05	104413.50	104639.80	**104000.00**	**104023.92**	104413.50	104639.80	**104095.00**	**104312.67**
wlp06	112169.50	112349.10	**111808.00**	**111915.25**	112169.50	112349.10	**112202.00**	**112326.45**
wlp07	163356.50	163489.60	**162760.00**	**162799.42**	163356.50	163489.60	**163125.00**	**163185.08**
wlp08	188851.50	189195.50	**188574.00**	**188582.00**	188851.50	189195.50	**188316.00**	**188461.67**
wlp09	219567.50	219879.00	**218849.00**	**219108.75**	219567.50	219879.00	**218663.00**	**218707.83**
wlp10	246586.50	246892.50	**245739.00**	**245836.67**	246586.50	246892.50	**245491.00**	**245664.75**
wlp11	290420.50	290620.40	**289364.00**	**289409.25**	290420.50	290620.40	**289066.00**	**289193.42**
wlp12	299146.00	299360.00	**297940.00**	**298086.25**	299146.00	299360.00	**297502.00**	**297637.33**
wlp13	316751.00	316879.20	**315340.00**	**315417.42**	316751.00	316879.20	**315225.00**	**315330.50**
wlp14	397169.00	397500.70	**395373.00**	**395618.00**	397169.00	397500.70	**395134.00**	**395531.67**
wlp15	460259.00	460909.00	**457843.00**	**458026.25**	460259.00	460909.00	**457438.00**	**457535.50**
wlp16	534789.50	535111.30	**531344.00**	**531795.75**	534789.50	535111.30	**530771.00**	**530895.25**
wlp17	566845.50	567518.70	**563344.00**	**563815.58**	566845.50	567518.70	**562509.00**	**562929.50**
wlp18	633046.50	633938.90	**628752.00**	**628847.25**	633046.50	633938.90	**627139.00**	**627389.83**
wlp19	752678.00	754195.30	**743934.00**	**744049.42**	752678.00	754195.30	**741398.00**	**741921.08**
wlp20	1000880.00	1004303.50	**976396.00**	**976783.92**	1000880.00	1004303.50	**971615.00**	**971997.58**
wlp21	**38128.00**	38237.80	38176.00	**38176.00**	38128.00	38237.80	**38067.00**	**38067.00**
wlp22	**74781.00**	75008.40	74899.00	**74912.50**	74781.00	75008.40	**74634.00**	**74710.67**
wlp23	**119622.50**	119806.90	119686.00	**119703.00**	119622.50	119806.90	**119412.00**	**119518.33**
wlp24	169886.00	170019.20	**169471.00**	**169648.50**	169886.00	170019.20	**169206.00**	**169234.33**
wlp25	231324.00	231536.90	**230477.00**	**230581.00**	231324.00	231536.90	**230453.00**	**230616.33**
wlp26	291883.50	292234.30	**291022.00**	**291088.67**	291883.50	292234.30	**290360.00**	**290467.83**
wlp27	391918.00	392156.80	**389970.00**	**390367.00**	391918.00	392156.80	**389331.00**	**389471.33**
wlp28	458944.50	459378.70	**456689.00**	**456805.58**	458944.50	459378.70	**455730.00**	**455877.75**
wlp29	597152.00	597880.80	**594279.00**	**594484.58**	597152.00	597880.80	**592512.00**	**592688.83**
wlp30	883828.00	884985.50	**871747.00**	**872360.58**	883828.00	884985.50	**868392.00**	**868714.33**
Avg	325860.70	326329.40	323296.80	323453.67	325860.70	326329.40	322623.67	322786.51

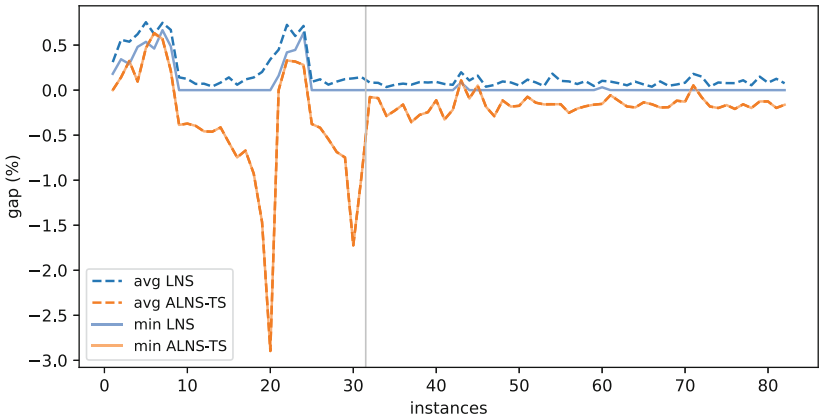


Fig. 1. GAP (%) per instance for LNS and ALNS-TS (long timeout). Negative values indicate solutions better than the BKS.

Table 3. Results for dataset `cflp-ci`, both timeouts.

Inst	$10\sqrt{m}$ seconds timeout				m seconds timeout			
	LNS		ALNS-TS		LNS		ALNS-TS	
	min	avg	min	avg	min	avg	min	avg
cflp-ci-00	418220.00	418769.90	**417946.00**	**418025.75**	417459.00	417808.80	**417026.00**	**417124.45**
cflp-ci-01	572832.00	575196.50	**558771.00**	**559844.50**	558300.00	558751.80	**556994.00**	**557481.91**
cflp-ci-02	445260.00	448563.50	**434599.00**	**434630.33**	434732.00	434882.70	**433483.00**	**433592.64**
cflp-ci-03	1507550.00	1513753.00	**1484520.00**	**1484716.75**	1477770.00	1478651.00	**1474278.00**	**1474904.55**
cflp-ci-04	1206190.00	1208104.00	**1204026.00**	**1204741.67**	1202900.00	1203768.00	**1200752.00**	**1201062.00**
cflp-ci-05	362702.00	363627.80	**357721.00**	**358103.92**	358155.00	358373.10	**356849.00**	**356992.82**
cflp-ci-06	153306.00	153663.10	**153096.00**	**153187.67**	153212.00	153349.10	**152779.00**	**152817.82**
cflp-ci-07	444237.00	445113.40	**439247.00**	**439818.00**	439747.00	440118.70	**438441.00**	**438663.82**
cflp-ci-08	695499.00	696752.50	**688982.00**	**689023.17**	688789.00	689417.00	**687619.00**	**688058.18**
cflp-ci-09	328371.00	329313.00	**325840.00**	**325944.25**	326288.00	326509.40	**325078.00**	**325239.73**
cflp-ci-10	604208.00	605182.70	**600782.00**	**601044.25**	601848.00	602204.70	**600272.00**	**600441.09**
cflp-ci-11	30746.00	30821.80	**30779.00**	**30779.00**	30772.00	30806.80	**30779.00**	**30779.00**
cflp-ci-12	454795.00	455467.90	**454375.00**	**454561.58**	454615.00	455101.90	**454193.00**	**454287.45**
cflp-ci-13	382532.00	383371.60	**382094.00**	**382268.67**	382412.00	383026.10	**381176.00**	**382154.36**
cflp-ci-14	1018520.00	1019821.00	**1015368.00**	**1015995.92**	1015270.00	1015657.00	**1012966.00**	**1013298.18**
cflp-ci-15	555115.00	556381.90	**550397.00**	**550855.42**	550969.00	551284.40	**549376.00**	**549747.09**
cflp-ci-16	878886.00	879700.80	**877093.00**	**877301.50**	876911.00	877754.10	**875826.00**	**876002.45**
cflp-ci-17	1218110.00	1219417.00	**1214177.00**	**1215260.58**	1212530.00	1213562.00	**1210262.00**	**1210754.45**
cflp-ci-18	893188.00	894583.40	**891207.00**	**891658.58**	891134.00	891594.60	**889582.00**	**890156.36**
cflp-ci-19	150444.00	150730.60	**150036.00**	**150163.33**	150326.00	150501.10	**150173.00**	**150195.82**
cflp-ci-20	1266230.00	1268089.00	**1260923.00**	**1261473.67**	1257690.00	1258765.00	**1255345.00**	**1255639.27**
cflp-ci-21	454159.00	454484.90	**452655.00**	**453088.25**	452760.00	452964.90	**451844.00**	**452034.73**
cflp-ci-22	**309532.00**	310162.10	309645.00	**309702.83**	309350.00	309914.20	**308590.00**	**308682.91**
cflp-ci-23	408511.00	408954.40	**406688.00**	**406854.58**	407059.00	407477.40	**406314.00**	**406405.82**
cflp-ci-24	465496.00	467897.30	**460252.00**	**460433.25**	460253.00	460698.20	**459030.00**	**459214.09**

(*continued*)

Table 3. (*continued*)

Inst	$10\sqrt{m}$ seconds timeout				m seconds timeout			
	LNS		ALNS-TS		LNS		ALNS-TS	
	min	avg	min	avg	min	avg	min	avg
cflp-ci-25	800345.00	803679.60	**780863.00**	**782003.42**	780002.00	780531.20	**778314.00**	**778484.00**
cflp-ci-26	689491.00	690836.70	**684720.00**	**685073.42**	683412.00	684091.60	**681708.00**	**682126.64**
cflp-ci-27	1430360.00	1436020.00	**1410422.00**	**1411582.58**	1405340.00	1405995.00	**1402442.00**	**1402920.27**
cflp-ci-28	463857.00	464817.80	**463336.00**	**463439.58**	464003.00	464335.60	**463063.00**	**463163.36**
cflp-ci-29	1370480.00	1374636.00	**1362198.00**	**1362425.92**	1356660.00	1357946.00	**1354323.00**	**1355330.27**
cflp-ci-30	573256.00	573932.80	**570948.00**	**571205.67**	571667.00	572104.40	**570262.00**	**570499.91**
cflp-ci-31	1202910.00	1207278.00	**1193317.00**	**1194103.33**	1191740.00	1192378.00	**1188694.00**	**1189288.36**
cflp-ci-32	482193.00	482863.10	**479186.00**	**479293.42**	478800.00	479249.60	**477774.00**	**477943.18**
cflp-ci-33	1071180.00	1073044.00	**1064393.00**	**1064523.33**	1060900.00	1061572.00	**1059286.00**	**1059571.82**
cflp-ci-34	1143900.00	1146552.00	**1138734.00**	**1139118.08**	1137760.00	1138195.00	**1135649.00**	**1135922.18**
cflp-ci-35	546574.00	549328.20	**530529.00**	**530698.33**	529424.00	529936.30	**528198.00**	**528557.00**
cflp-ci-36	548831.00	549771.90	**546476.00**	**546576.58**	546761.00	547024.40	**545255.00**	**545733.36**
cflp-ci-37	1275540.00	1276290.00	**1272054.00**	**1272499.58**	1269810.00	1270609.00	**1268299.00**	**1268602.09**
cflp-ci-38	1044810.00	1045944.00	**1042490.00**	**1043008.25**	1042340.00	1043200.00	**1040646.00**	**1041029.09**
cflp-ci-39	190422.00	190673.10	**189721.00**	**190121.92**	190241.00	190584.20	**190124.00**	**190324.45**
cflp-ci-40	630056.00	630750.30	**627629.00**	**628032.50**	627341.00	628284.50	**626632.00**	**626912.73**
cflp-ci-41	1308560.00	1313895.00	**1284233.00**	**1284616.00**	1279750.00	1280235.00	**1277021.00**	**1277684.82**
cflp-ci-42	417079.00	417824.10	**416317.00**	**416397.75**	416525.00	416879.00	**415698.00**	**415913.64**
cflp-ci-43	501463.00	503412.80	**497322.00**	**497380.67**	496873.00	497260.80	**495732.00**	**495922.00**
cflp-ci-44	759045.00	763009.50	**748200.00**	**749311.42**	748310.00	748891.20	**746744.00**	**746887.55**
cflp-ci-45	714632.00	715128.10	**711231.00**	**711814.42**	711314.00	712092.30	**709883.00**	**710339.45**
cflp-ci-46	626501.00	628046.60	**622111.00**	**622520.25**	621462.00	621798.80	**620142.00**	**620312.27**
cflp-ci-47	405474.00	406156.90	**405102.00**	**405343.58**	405218.00	405824.70	**404364.00**	**404707.82**
cflp-ci-48	768781.00	769566.70	**767034.00**	**767180.00**	766996.00	767625.10	**765513.00**	**765822.73**
cflp-ci-49	300901.00	301208.90	**299875.00**	**300108.67**	300328.00	300705.40	**299735.00**	**299811.91**
Avg	689825.60	691451.78	684593.20	684957.12	683884.56	684405.82	682501.36	682790.84

5.4 Statistical Significance Analysis

While average and best-case results provide important evidence of algorithmic performance, they are insufficient to determine whether observed improvements are due to stochastic variability. To assess whether the gains achieved by ALNS-TS over the baseline LNS are statistically significant, we conduct a hypothesis testing analysis based on paired non-parametric tests.

Each algorithm was executed independently ten times per instance, yielding stochastic solution distributions with non-normal characteristics. These characteristics motivate the use of non-parametric paired tests, which do not rely on distributional assumptions and are recommended in the evaluation of metaheuristic algorithms.

We employ the Wilcoxon signed-rank test [5,9] to compare LNS and ALNS-TS under the short timeout ($10\sqrt{m}$ seconds). For each instance, the average solution cost over the ten runs is computed for both algorithms, and the test is applied to the paired per-instance differences. The null and alternative hypotheses are defined as

$$H_0 : \text{median}(C_{\text{LNS}} - C_{\text{ALNS-TS}}) = 0, \quad H_1 : \text{median}(C_{\text{LNS}} - C_{\text{ALNS-TS}}) > 0,$$

where C denotes the solution cost. The alternative hypothesis reflects the expected directional improvement of ALNS-TS, which aims to achieve lower costs through adaptive operator selection. All tests are conducted at a significance level of $\alpha = 0.05$.

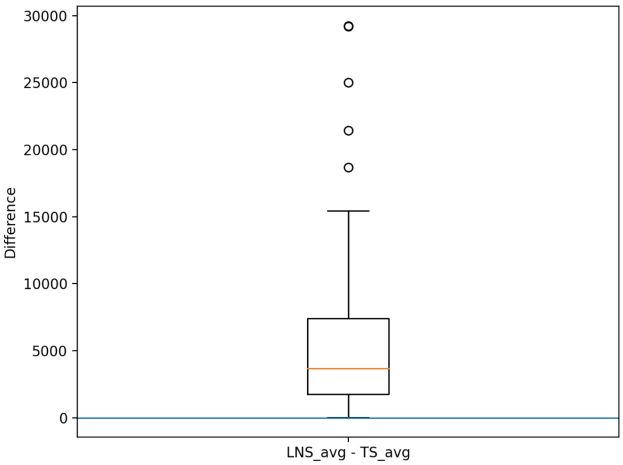


Fig. 2. Paired differences in per-instance average costs for `cflp-ci` ($C_{\text{LNS}} - C_{\text{ALNS-TS}}$). Positive values indicate ALNS-TS is better.

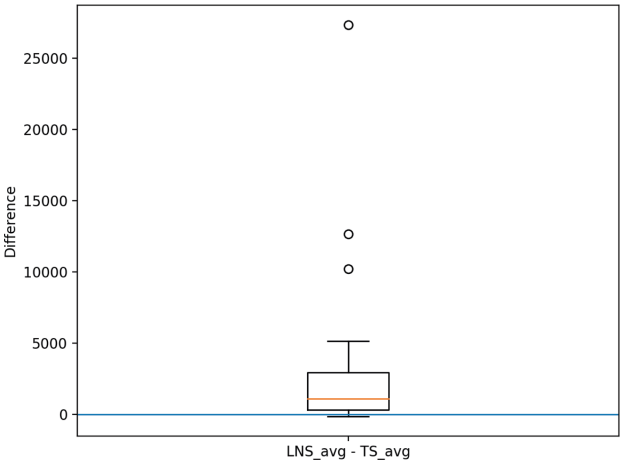


Fig. 3. Paired differences in per-instance average costs for `wlp` ($C_{\text{LNS}} - C_{\text{ALNS-TS}}$). Positive values indicate ALNS-TS is better.

Figures 2 and 3 summarize the paired differences in per-instance average costs, $C_{\text{LNS}} - C_{\text{ALNS-TS}}$. Positive values indicate instances for which ALNS-TS outperforms LNS.

Across both instance sets, the distributions of paired differences are predominantly positive, with median values clearly above zero. This indicates that, for most instances, ALNS-TS consistently achieves lower average costs than LNS. Although a small number of instances exhibit near-zero differences, these cases are rare and of limited magnitude when compared to the overall cost scale.

The Wilcoxon signed-rank test confirms that the observed improvements of ALNS-TS over LNS are statistically significant for both instance sets. In particular, the *p-value* the test rejects the null hypothesis at the 5% significance level, indicating that the median per-instance average cost achieved by ALNS-TS is lower than that of LNS.

For the `wlp` instances, the p-values are $p = 9.64 \times 10^{-8}$ for the minimum-cost comparison and $p = 4.66 \times 10^{-9}$ for the average-cost comparison. For the larger `cflp-ci` instances, the corresponding p-values are even smaller, with $p = 3.26 \times 10^{-10}$ for minimum costs and $p = 2.57 \times 10^{-10}$ for average costs.

These small p-values provide statistical evidence that the median per-instance solution cost achieved by ALNS-TS is lower than that of LNS, ruling out stochastic variability as an explanation for the observed performance differences.

The results reflect a reliable improvement induced by the learning-based operator selection mechanism. By adapting operator choices online and in a context-aware manner, Thompson Sampling enables a more effective balance between intensification and diversification, resulting in both lower average costs and improved stability across independent runs.

Overall, the statistical analysis provides evidence that integrating Thompson Sampling into the ALNS framework yields a statistically significant enhancement over the LNS approach.

6 Conclusions

This paper introduced ALNS-TS, an Adaptive Large Neighborhood Search framework augmented with a contextual Thompson Sampling mechanism for operator selection, targeting the Multi-Source Capacitated Facility Location Problem with Customer Incompatibilities (MS-CFLP-CI). The proposed approach preserves the core structure of the LNS algorithm, proposing a new conflict-aware operator and replacing traditional segment-based operator adaptation with an online, context-aware selection strategy.

Computational results on the `wlp` and `cflp-ci` benchmark sets show that ALNS-TS improves both best-case and average solution quality over the current state-of-the-art LNS, with statistically significant gains across both benchmarks. These gains are observed across different instance sizes and time budgets, indicating that the proposed learning mechanism is effective in the search process.

The methodological contribution of this work has two components with different scopes. The contextual Thompson Sampling mechanism for operator selection is problem-agnostic: it relies only on a generic search-state feature vector

(progress, temperature, stagnation, acceptance signals, and structural ratios) and on a Bayesian linear model per arm, which can be transferred to any ALNS-based solver with multiple destroy/repair operators. In particular, vehicle routing, scheduling, and network design problems for which ALNS is a standard approach could benefit directly from the same selection scheme, possibly augmented with a few problem-specific context features. In contrast, the conflict-aware CCF operator is specifically tailored to MS-CFLP-CI, as it exploits the incompatibility graph structure of the problem; analogous problem-aware operators would need to be designed for other domains, but the integration pattern with contextual Thompson Sampling remains the same.

Several directions for future research emerge from this work. First, richer context representations and alternative reward definitions could be explored to further improve credit assignment and responsiveness to search dynamics. Second, the proposed framework extends to settings with multiple repair operators, including hybrid exact–heuristic repair strategies.

Overall, this paper demonstrates that embedding contextual Thompson Sampling into the ALNS framework yields a robust and statistically significant improvement over classical LNS approaches, highlighting the potential of learning-based operator selection as a general paradigm for complex combinatorial optimization problems.

References

1. Abeille, M., Lazaric, A.: Linear Thompson sampling revisited. Artif. Intel. Stat. (2017)
2. Agrawal, S., Goyal, N.: Thompson sampling for contextual bandits with linear payoffs. In: Proceedings of the 30th International Conference on Machine Learning, pp. 127–135. ICML (2013)
3. Ceschia, S., Schaerf, A.: Multi-neighborhood simulated annealing for the capacitated facility location problem with customer incompatibilities. Comput. Ind. Eng. **188**, 109858 (2024). https://doi.org/10.1016/j.cie.2023.109858
4. Gjergji, I., Kletzander, L., Musliu, N., Schaerf, A.: Large neighborhood search for capacitated facility location with customer incompatibilities. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '25 (2025). https://doi.org/10.1145/3712256.3726355
5. Hettmansperger, T.P., McKean, J.W.: Robust Nonparametric Statistical Methods. CRC Press (2010)
6. Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press (2020)
7. Maia, M.R.H., et al.: Metaheuristic techniques for the capacitated facility location problem with customer incompatibilities. Soft. Comput. **27**, 4685–4698 (2023). https://doi.org/10.1007/s00500-022-07600-z
8. Osband, I., Russo, D., Van Roy, B.: More efficient reinforcement learning via posterior sampling. In: Advances in Neural Information Processing Systems (2013)
9. Pratt, J.W.: Remarks on zeros and ties in the Wilcoxon signed rank procedures. J. Am. Stat. Assoc. **54**(287), 655–667 (1959)
10. Revelle, C.S., Laporte, G.: The plant location problem: new models and research prospects. Oper. Res. **44**(6), 864–874 (1996)

11. Røpke, S., Pisinger, D.: An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. Transp. Sci. **40**(4), 455–472 (2006). https://doi.org/10.1287/trsc.1050.0135
12. Russo, D., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z.: A tutorial on Thompson sampling. Found. Trends Mach. Learn. **11**(1), 1–96 (2018)
13. Thompson, W.R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika **25**(3–4), 285–294 (1933)

MSA Strategies for End-of-Line Automated Test Equipment with Non-replicable Measurements

Eduardo Santos(✉), Eusébio Nunes, and Cristina Rodrigues

University of Minho, 4710-057 Braga, Portugal
ID11155@uminho.pt

Abstract. Measurement data quality is the fundamental pillar for statistical process control and reliable decision-making in modern industry. However, a persistent deficiency in scholarly literature regarding measurement system analysis (MSA) for complex electronic devices is observed. This article addresses MSA in automated industrial environments, focusing on automated test equipment (ATE) used in End-of-Line (EOL) testing in the automotive industry, where non-replicable measurements challenge traditional methods such as Gage R&R (GRR). The main objective is to propose and validate methodological strategies capable of evaluating the variability of the measurement system in these contexts, minimizing decision errors (Type I and II). To this end, an approach based on Expanded GRR combined with ANOVA was adopted, allowing the variance to be decomposed into different sources, such as parts, benches, nests, and repeatability error. The experimental study involved 19 devices tested in multiple configurations over 13 weeks, with statistical analysis performed using Minitab. The results revealed that the measurement system has a total GRR of 36.59%, considered inadequate, with significant contributions from repeatability (13.97%) and reproducibility (22.62%), in addition to statistically significant differences between benches, nests, and their interactions. It is concluded that classical methods are insufficient for non-replicable measurements, making it necessary to search for more robust models such as the Expanded GRR with ANOVA. The study reinforces the importance of considering reproducibility in ATE, improving equipment calibration and maintenance, and developing future research that relates measurement variability to product quality and industrial decisions.

Keyword: MSA · ATE · Non-replicable measurements

1 Introduction

In industries such as semiconductors and automotive electronics, Measurement System Analysis (MSA) [7] plays a critical role in quantifying measurement error and ensuring that variability introduced by the system does not compromise the correct functional classification of products. Inadequate characterization of the measurement system can lead to Type I errors - the improper rejection of conforming devices, directly impacting yield - or Type II errors - the improper acceptance nonconforming devices, with potential consequences for reliability and user safety [6].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 396–412, 2027.
https://doi.org/10.1007/978-3-032-30494-0_25

Concurrently, the quality of measurement data is widely recognized as a determining factor for the robustness of technical, economic, and strategic decisions [20, 27] gaining particular relevance in regulated sectors like the automotive industry. In this context, Integrated Device Manufacturers (IDM) are urged by customers and auditors to implement systematic and increasingly reliable methods for the classification of each Device Under Test (DUT), in compliance with regulatory and contractual requirements [3].

The centrality of metrological quality is reflected in international standards such as ISO 9001 [14], which requires the suitability of monitoring and measurement resources, and ISO 10012 [15], which establishes specific requirements for measurement management systems. Complementarily, ISO 5725 [13], the GUM (ISO/IEC Guide 98-3) [11], and ISO/IEC 17025 [17] provide the technical framework for evaluating precision, uncertainty, and validity of measurement methods and results. Together, these standards consolidate the requirement that measurement systems be "fit for purpose" and maintain their capability over time.

1.1 Classical MSA Approaches

MSA studies, particularly GRR, are based on the fundamental assumption of the replicability of measurements on the same object under identical conditions.

According to the Automotive Industry Action Group (AIAG) [7], the objective of MSA is to evaluate whether a measurement system consistently distinguishes conforming from non-conforming devices, thereby minimizing incorrect decisions associated with system variability. This methodology is widely applied in instrument validation, production process qualification, and structured quality improvement initiatives, allowing for the separation of inherent part-to-part variability from the variability introduced by the measurement system [4].

In mass production environments, the total error of the measurement system can be decomposed into components such as repeatability, reproducibility, number of distinct categories (ndc), bias, and temporal stability, as recommended by the AIAG manual [7]. Among these, repeatability and reproducibility - usually evaluated through GRR studies - are particularly relevant for the continuous improvement of process performance [6].

One of the main advantages of MSA, as established by the AIAG, lies in avoiding the complexity associated with full measurement uncertainty modeling [7]. Instead, it focuses on the practical evaluation of measurement system capability relative to the functional requirements of the process and product. In this sense, determining the capability of a measurement system is not only an essential element of quality and process improvement initiatives [4] but also a critical step that must precede any effort to improve the design or production process, as recommended by industrial best practices and standards [1].

It is important to note that despite its widespread use, a misconception persists in industrial practice that GRR studies directly evaluate measurement uncertainty. This interpretation is incorrect, as GRR is limited to analyzing the variability introduced by the measurement system and does not provide information regarding the interval within which the true value of the measurand is expected to lie, through calibration [22]. Thus, in accordance with ISO 10012 [15], ISO/IEC 17025 [17], and the principles of metrology established in the Guide to the Expression of Uncertainty in Measurement

(GUM), the calibration of measuring instruments must be viewed as a necessary and complementary activity to MSA, essential for ensuring metrological traceability and the proper evaluation of measurement uncertainty.

Classical MSA studies reveal significant limitations in industrial environments characterized by high automation, ATE, and complex hierarchical architectures.

1.2 MSA in Automatic Test Equipment

An ATE is a sophisticated hardware and software infrastructure designed to perform automatic measurements and diagnostics on components, sub-assemblies, or DUT. In the context of the automotive electronics industry, these systems are crucial for ensuring that products meet the rigorous performance and safety characteristics required by customers and quality auditors, such as those governed by the IATF 16949 [10] standard. In practice, the ATE replaces the human appraiser with a fully automated setup capable of performing parallel tests on multiple components more efficiently and at a lower cost. However, these systems introduce additional sources of variability, such as differences between various test sites and benches used [3].

The reliability of an ATE in the automotive electronics sector depends on the statistical equivalence of the different benches, meaning that different benches of the same ATE must produce identical results for the same device under test, regardless of the test bench, nest, time, or geographic location. The adequacy of this equipment is verified through MSA study methodologies, where the ATE is evaluated as an important component of the measurement chain to minimize Type I and Type II errors.

ATE introduces multiple sources of variability whose interaction frequently violates the assumptions of independence and homogeneity of variance inherent in standard MSA models [1]. This issue is exacerbated in non-replicable measurement processes.

This paper addresses this theme, building upon the study by Santos et al. [25], whose results highlighted the need for more in-depth MSA studies on ATE for automotive electronic systems. More specifically, this study aims to evaluate the measurement capability of an ATE used in End-of-Line (EOL) testing of automotive electronic devices, considering the differences between test benches and between nests within the same bench [3].

The remaining sections of this article are organized as follows: Sect. 2 presents EOL testing for automotive electronic devices, highlighting the framework and objectives of these tests, their main characteristics, and the non-replicable nature of many of the measurements obtained. Section 3 covers the main Experimental Design approaches in MSA studies, with special emphasis on those suitable for non-replicable measurements. Section 4 presents the methodology used in the MSA study described in detail in Sect. 5. Finally, Sect. 6 presents the main conclusions of this study.

2 EOL Testing in Automotive Electronic Devices

EOL testing in automotive electronic systems/devices comprises the set of tests performed at the final stage of the assembly process, immediately prior to camera measurement inspection (CMI), packaging, and shipping to the customer for vehicle integration.

EOL serves as the final technical barrier for verifying the functional, electrical, and, in many cases, communication compliance of automotive electronic modules, playing a critical role in ensuring the quality, safety, and reliability of the electronic device.

2.1 EOL Objectives and Tests Coverage

In the automotive industry, quality requirements are particularly demanding, both due to the regulatory framework (e.g., IATF 16949) [10] and the direct implications for road safety and brand reputation. Therefore, the primary objectives of EOL testing are:

- To verify the functional compliance of the final product.: activation of outputs (actuators, relays, drivers); reading of inputs (simulated sensors); validation of control algorithms; boundary condition response tests;
- To ensure that electrical and performance parameters are within specified tolerances: measurement of voltages, currents, and power consumption in different operating modes; verification of analog and digital signals; supply integrity and short-circuit protection tests;
- To confirm the correct programming, coding, and configuration of the software included in the final product;
- To guarantee the individual traceability of the tested product.

EOL tests can also include:

- Communication tests: verification of automotive protocols such as CAN, LIN, or Ethernet; validation of identifiers, firmware, and coding;
- Functional safety tests: in critical modules (e.g., braking systems), specific tests associated with functional safety requirements (ISO 26262) [16] may be performed.

Following these tests, the EOL system performs the device parameterization to standard values, preparing it for future pairing with the vehicle.

2.2 Main Characteristics of EOL Testing

EOL tests in automotive electronics present a set of distinctive characteristics, such as:

- High degree of automation between device and test system - with multiple channels and high hardware and software integration;
- Reduced cycle time – there is significant pressure to minimize test duration without compromising coverage;
- High functional coverage – unlike purely parametric tests, EOL tests tend to be highly functional, seeking to reproduce real operating scenarios;
- Full traceability – each tested device is typically associated with a unique serial number, time stamp, and a complete record of test results;
- Binary classification with strict criteria – EOL tests generally results in a pass/fail decision.

2.3 Non-replicable Measurements in Eletronic Devices

According to De Mast, J. and Trip, A. [18], a measurement is considered non-replicable when it violates one of two principles: (i) Temporal stability: the true value of the object changes over time; (ii) Robustness against measurement: the object is destroyed or significantly altered during the test.

Under these circumstances, the empirical separation between measurement system variability and the intrinsic variability of the object or measured characteristic becomes impractical, compromising the fundamental assumptions of conventional MSA methods. Despite the recognition of these limitations, a methodological gap persists regarding formally structured statistical frameworks for complex ATEs and non-replicable measurements [3].

Electronic circuits, due to their architecture, exhibit real variability, under controlled and identical conditions [29]. The variability can be classified at the hardware level, including manufacturing variations, thermal effects, electrical fluctuations, aging phenomena, and layout parasitic [21]. On hardware, variability derives from jitter and discrete components (i.e., resistors, capacitors, inductors) [23, 28]. Furthermore, manufacturing variations dictated by material composition and fabrication techniques also introduce variations [5]. Software also introduces variability: temporal (latency, task priority, interrupts); state (buffers, queues); environmental (drivers, libraries, OS versions, CPU load, and memory availability); as well as algorithmic and communication factors (bus latency, packet loss, and communication jitter). Other factors relate to the interaction between components, such as processing frequency, CPU load, and disk/memory read/write operations [19]. These are, in turn, associated with the system state (including switching between active and sleep modes of various sub-circuits) and interference from base software, such as operating system task scheduling-which includes instructions, routines, various software modules, and several network connections that also consume energy [8]. Both the circuit and ambient temperature also influence the results obtained [28].

In these cases, as in many other advanced industrial contexts, the fundamental assumption of classical MSA methods - the replicability of measurements on the same object under identical conditions - is not verifiable. Consequently, it is necessary to employ alternative MSA strategies that preserve statistical validity and metrological significance.

3 Experimental Designs in MSA Studies

The literature on MSA studies for non-replicable measurements suggests various methods or experimental strategies to isolate the different sources of variability within a measurement system, namely: Crossed designs [9, 22, 27]; Nested designs [9, 22, 27]; Expanded designs [22, 27]; Latin Square designs [26]; Mixed-Effects Models [1, 27]; and Type Gage 1 studies [22].

Resume. In non-replicable measurements, MSA requires a paradigmatic shift away from classical approaches. The choice of the most appropriate methodology depends on the system architecture, experimental constraints, and the objective of the analysis.

3.1 Crossed Design

This is the most common approach for evaluating measurement system variability through GRR studies. Every part is measured by every operator, and each operator-part combination is repeated several times, allowing for a fully crossed structure between factors. It allows the total variability to be decomposed into repeatability (equipment variation), reproducibility (variation between operators), and part-to-part variability. However, the disadvantages are (i) the need of a high number of measurements, which can involve significant cost and time; and (ii) assumption of stable experimental conditions and the absence of operator learning or fatigue effects. Several authors demonstrate usability for non-replicable measurements [9].

3.2 Nested Design

Nested designs constitute one of the most used approaches for non-replicable measurements. This model is essential when measurements are destructive or when parts cannot be measured more than once by the same operator. Each part is exclusive to one operator or condition, ensuring that two operators never measure the same device. However, the nested design cannot be used to estimate the reproducibility, and no additional factors can be added.

3.3 Expanded Design

Expanded designs are an extension of crossed and nested GRR designs, aimed at accommodating greater experimental complexity and it incorporates additional fixed or random factors (in addition to parts and operators), such as multiple pieces of equipment, environmental conditions, production batches, or operational levels, allowing for a more comprehensive assessment of measurement system performance in real-world contexts. Among the primary advantages of these experimental designs is their greater representativeness of the actual system, allowing for the capture of interactions between multiple sources of variability. However, their implementation is more complex, requiring more rigorous experimental planning, a larger volume of data, and more advanced statistical skills. Nevertheless, expanded designs do not, by themselves, solve the problem of the absence of replication.

3.4 Other MSA Designs

Latin Square designs provide an efficient, robust, and flexible method for controlling two sources of variability in non-replicable measurements, although they are restricted by equal factor levels and the assumption of no interactions [26]. Linear mixed-effects models offer a more versatile alternative by treating measurement system components as random effects to decompose variance in complex, unbalanced scenarios, despite requiring advanced statistical expertise [1, 27]. Type 1 Gage Study is also an alternative to classical MSA, for quick check of bias and repeatability of a single device through Cg and Cgk, used in ATE [22].

4 Methodology

Given the specific characteristics of ATE systems used in automated industrial environments - namely the high degree of automation and the lack of direct operator intervention - this MSA study adopted an approach based on the expanded GRR method. The implementation steps of this methodology are presented in the following sections.

4.1 Statistical Model

For data analysis, a random-effects ANOVA model was considered, allowing for the estimation of the various variance components (factors) associated with the measurement system. Establishing the factors to be included in the study is a fundamental aspect. Examples of factors included in the analysis are:

- Part (P): represents the intrinsic variability between the devices under test;
- Test Station (S): different test benches (ATE);
- Nest (N): physical contact interfaces with the bench (ATE);
- Repetition (ε): random error associated with the measurement process.

The general model may be expressed as:

$$Y_{ijkl} = \mu + P_i + S_j + N_k + \epsilon_{ijkl} \tag{1}$$

where:

Y_{ijkl} represents the observed value

μ is the global mean

P_i, S_j, N_k correspond to the random or fixed factors effect

ϵ_{ijkl} represents the repeatability error

The variance components are estimated based on this model, allowing for the quantification of the relative contribution of each source of variation.

4.2 Experimental Procedure

The experimental procedure was carried out in the following stages:

1. Selection of the variant based on statistical representativeness criteria;
2. Execution of the measurements in accordance with the established sampling plan;
3. Automatic data recording by the ATE.

4.3 Estimation of Variance Components

Total variance decomposition was performed based on Expanded and ANOVA analysis, allowing for the estimation of repeatability variance; variance associated with benches; variance associated with nests; bench-nest interaction variance; and part-to-part variance. Based on these components, the classical MSA indicators were calculated:

- Percentage of GRR (%GRR);
- Percentage of Repeatability;
- Percentage of Reproducibility;
- Number of Distinct Categories (ndc).

4.4 Acceptance Criteria

The suitability of the measurement system was evaluated based on criteria widely accepted in both literature and industrial practice:

- %GRR $< 10\%$: Suitable measurement system;
- $10\% \leq$ %GRR $\leq 30\%$: Acceptable system, depending on the importance of the application, the cost of the measurement device, or the cost of repair;
- %GRR $> 30\%$: Inadequate system; every effort should be made to improve it.

5 Case Study

This section presents the MSA study conducted on the ATE at the EOL assembly lines for automotive electronic products. These products are manufactured by a company in Portugal belonging to a multinational group. Currently, the company produces its devices according to defect acceptance criteria based on Class 3, with reference to the IPC Standard: Acceptability for Electronic Assemblies (IPC-A-610G) [12].

The analysis was processed through Minitab® Statistical Software.

5.1 ATE Characterization

ATE evaluates the device once it is fully assembled and encapsulated, in debug mode, under conditions as close as possible to real-world use, integrated at the end of the assembly/production line (see Fig. 1). The line has a total of eleven test benches with two nests each (two distinct physical interfaces). Each bench features a rack equipped with different measurement devices, allowing for the evaluation of a diverse set of functionalities and variables.

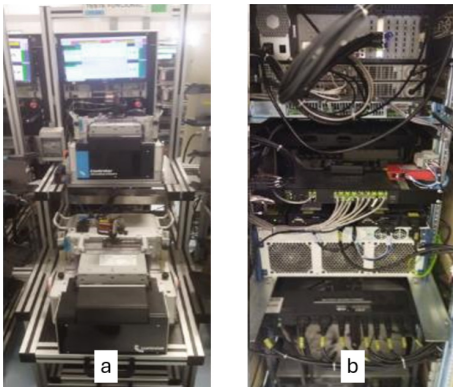


Fig. 1. Automatic EOL test bench, front (a) and rear view (b)

5.2 DUT Characterization

The product used in this study, hereafter referred to as the device (Fig. 2), is an infotainment system that combines informational functionalities (e.g., receiving signals from

the vehicle's peripherals) and entertainment (e.g., AM/FM, USB and Bluetooth protocols). Additionally, a group of Critical-to-Quality (CTQ) characteristics is established and measured, based on customer and supplier requirements.

Fig. 2. Electronic device tested in EOL

All products are functionally tested in ATE. Each functional characteristic corresponds to a specific number of numerical and attribute variables, with a final Pass/Fail. Furthermore, each type of measurement variable corresponds to a specific piece of equipment installed in the rack. As an example:

- For variables reflecting current consumption, the measurement equipment is a Digital Multimeter (DMM), common to both nests;
- For variables reflecting audio signal measurements or signal-to-noise ratios (FM, AM, DAB, etc.), the measurement equipment used is a Digital Signal Analyzer (DSA).

5.3 Adopted MSA Method in the Automotive Company

The method currently used by the automotive company to conduct MSA studies on ATE involves collecting a sample of ten consecutive parts of the same variant from the production process during a given shift.

The frequency of these studies is determined by customer requests or significant changes to the process or equipment. The statistical method employed is the Range method for a quick assessment and convenience.

The characteristics of each sample element are evaluated through four test runs on the same nest within one ATE. With the current MSA approach, the company assumes there are no differences between the independent ATEs and, consequently, considers reproducibility to be ignored. However, according to several authors [3, 24] significant differences may indeed exist between the ATEs. Furthermore, parts are collected sequentially (convenience sampling); therefore, the study fails to capture the real process variability. This highlights the need for a more detailed and methodologically demanding MSA study.

In this regard, the present paper acts as a continuation of the work by Santos, E. et al. [25], where results showed inconclusive trends and deviations between ATEs. That evaluation underscored the necessity for more in-depth MSA research, specifically applied to electronic systems.

5.4 Samples Preparation and Collection

On this device, true values change over time, therefore non-replicable measurements methods apply. The first step involves reducing measurement instability as much as possible. Mikulová, P et al. [20] address several general scenarios for this. Following the authors' recommendations, a homogeneous batch was selected - specifically, the same variant for the study - so that the devices could be considered "identical" for the analysis. Same variant was chosen, which consists of an electronic device with a set of identical attributes (same software version, hardware, and quantity and type of components), sharing the same unique identification name and ID number.

Sampling Plan. The study involved EOL testing on one of the production lines previously identified by Santos, E. et al. [25] as the most unstable. Sample collection took place over 13 non-consecutive weeks, and the respective tests were conducted throughout this same period. Benches with ID 1, 3, 5, and 7 were used for this study due to shift time constraints.

Since the line was not in production during the third shift (overnight), this time slot was utilized for the testing. An operator was assigned exclusively to perform the tests, representing a deliberate investment in extra time and cost.

The experimental plan was defined to ensure the representativity of process variability and the statistical robustness of the analysis. N devices were selected, assuring statistically the inclusion of some near the specification limits. A total of 19 devices for a given variant were subjected to the MSA test and were tested three times in each bench-nest combination. The sampling plan followed in this study is presented in the flowchart in Fig. 3.

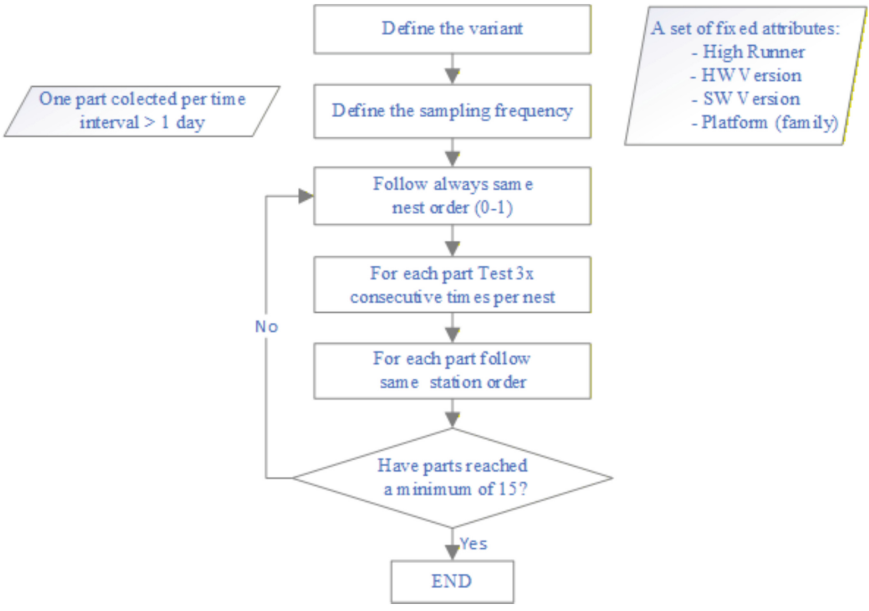


Fig. 3. Sampling plan

Part-to-Part Variation - Unique Parts Count. If the parts are stable and the randomness of the measurement error is homogeneous, a GRR can be performed [1].

Fourteen device variables were analyzed to select the most suitable ones for the study.

To ensure that the observed total variability was in greater proportion, due to actual part-to-part variation rather than random measurement error, variables that displayed a significant percentage of part-to-part variation were sought - specifically, those with a P-value equal to or less than 0.005 (very close to zero).

Additionally, the count of unique parts under this condition, for each variable, had to exceed 50% of the total tested samples. To achieve this, a Factorial ANOVA method (General Linear Model) was employed.

It was observed that five variables out of fourteen met this condition. Under these conditions, there is a high probability that repeatability and reproducibility represent a lower relative percentage than the part-to-part variation.

Normal Distribution. Before proceeding with the ANOVA calculations, the normality of the data for each selected variable was verified. This is a fundamental assumption for the validity of the GRR results, according to AIAG MSA [7].

Out of the five variables, it was found that only one, "102411 A2B+ Boost", fits a normal distribution at a 10% significance level, as shown in Fig. 4.

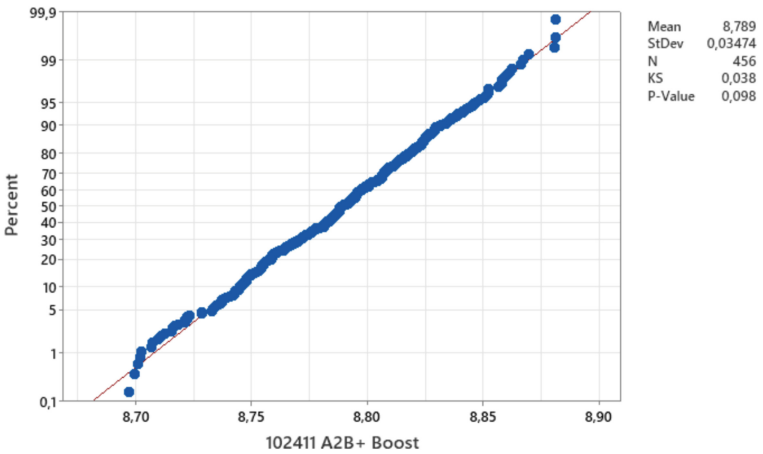


Fig. 4. Probability plot

It is worth noting that the CTQ customer variable, "Check Standby Current", follows a normal distribution; however, it exhibits very low part-to-part instability.

ANOVA – Time and Test Order Test. To determine whether significant differences exist in device measurements over time, a factorial ANOVA was performed. For the selected variable 102411 A2B + Boost, it was observed that significant differences occur over time between samples (measured week after week), and that significant differences exist within each sample according to the measurement sequence (test order).

Characterization of the Selected Variable. With the above criteria met, the variable selected for the study is "102411 A2B + Boost". Digital audio is measured using various parameters, specifically through the audio signal and current consumption.

This variable is one of the parameters used to measure digital audio; its function is to test the digital audio output via current consumption, measured in Volt. The signal is sent to the device ("unit" in the picture) through an individual generator installed in each nest, which then processes it and returns it to one of the common measurement devices on the bench, the DMM - as shown in Fig. 5.

The specification limits are LSL = 8.000000 V and USL = 9.000000 V.

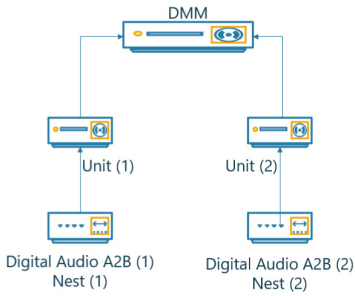


Fig. 5. Bench and Nest layout

5.5 MSA Expanded

Below, the application of the Expanded GRR and ANOVA method is highlighted. Since each ATE is independent and the nests share measurement equipment, it is considered an "operator".

5.6 Analysis Results

The ndc for this variable is equal to 1. According to the AIAG MSA manual [7], this indicates that the measurement system may be unable to identify distinct categories of parts.

However, according to VDA 5 [30], this parameter can be disregarded if the equipment resolution is at most 5% of the tolerance, by ensuring the quality of the readings. With adequate resolution, the determination of the ndc becomes dispensable, provided that the parts in the GRR have different measurements. Otherwise, the measurement system will fail to detect differences between parts [2]. In this study, the resolution of the equipment is 0.001% of the tolerance unit vector (0.1). Therefore, the ndc parameter can be safely ignored.

ANOVA study shows significant differences between devices ("Recoded Barcode"), test benches, nests and at interaction of Bench*Nest, as shown in Table 1.

Table 1. ANOVA Table with All Terms

Source	DF	Seq SS	Adj SS	Adj MS	F-Value	P-Value
Recoded Barcode	18	0.34503	0.34503	0.019168	320.99	0.000
Bench	3	0.08592	0.08592	0.028641	479.61	0.000
Nest	1	0.04280	0.04280	0.042797	716.67	0.000
Bench*Nest	3	0.04980	0.04980	0.016600	277.98	0.000
Repeatability	430	0.02568	0.02568	0.000060		
Total	455	0.54923				

α to remove interaction term = 0,05.

In the Components of Variation graph, %Contribution indicates the share of each source in the total variance and %Study Var shows the contribution of each source to the study's variability measured on the standard deviation scale. Both percentages sum to a cumulative 100%. See Fig. 6.

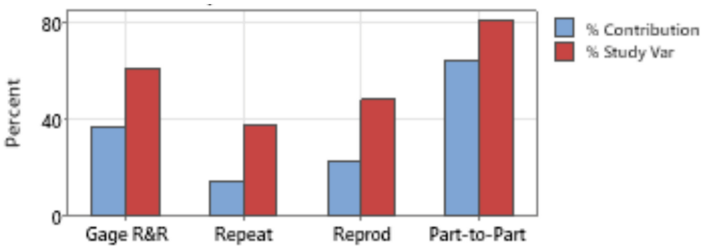


Fig. 6. Component of Variation

In Table 2, it is observed that the part-to-part component dominates, accounting for 63.41% of the variance between devices.

Total GRR accounts for 36.59% of the variance, resulting from the sum of Repeatability with 13.97% and Reproducibility with 22.62%.

Table 2. Variance Components result

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	0.0004566	36.59
Repeatability	0.0001743	13.97
Reproducibility	0.0002823	22.62
Bench	0.0001884	15.10
Nest	0.0000939	7.52

(*continued*)

Table 2. (*continued*)

Source	VarComp	%Contribution (of VarComp)
Part-To-Part	0.0007914	63.41
Recoded Barcode	0.0007914	63.41
Total Variation	0.0012480	100.00

The Sample Range chart, Fig. 7, represents the variability within each subgroup or part (max-min), where small and consistent ranges indicate good repeatability. This chart helps analyze and pinpoint specific causes of significant measurement system variation; it is observed that Bench 5 shows significant differences compared to the other benches. This suggests equipment instability on measurement capability.

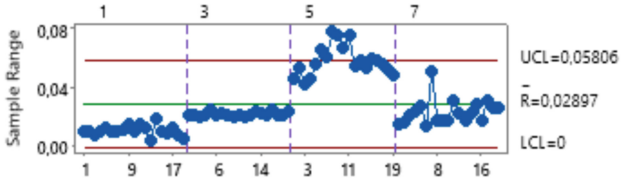


Fig. 7. R-Chart result

The Sample Mean chart, Fig. 8, shows the variability between different subgroups or parts and helps evaluate process trends (drift) and outliers that affect the measurement system assessment. A large part-to-part dispersion is observed (indicating a healthy difference between devices).

It is also observed that the process is out of control, regardless of the specification limits. The control limits are based on 3*standard deviation (STDEV = 0.0347); therefore, there is a significant probability of the population exceeding the specification limits.

In sum, the chart highlights a significant low systematic bias among three out of four benches (measuring similar levels) but may require verification of Bench seven (with higher average) and a review of the special causes for the points above the limits.

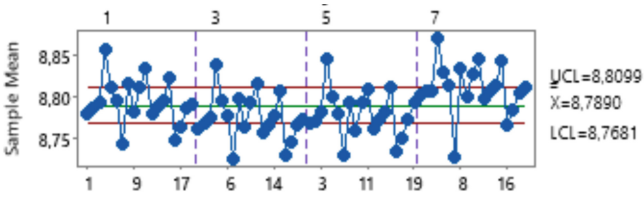


Fig. 8. X-Bar Chart result

In the Interaction Plot between Bench and Nest (Fig. 9), a significant difference is observed in the means within subgroups (max-min) - specifically between Nests - on Benches three and five.

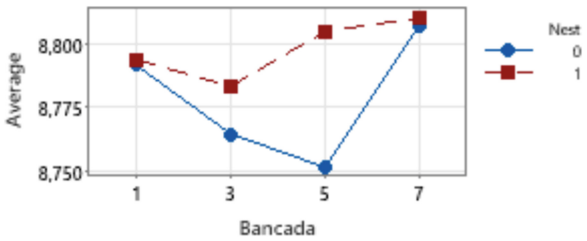


Fig. 9. Bench (*Bancada*)*Nest Interaction

6 Results Discussion

The results suggest that the measurement system shows inconsistencies which require attention and improvement. In benches three and five, disparity between nests must be analyzed as it raises suspicions regarding the consistency of the signal generators, or about the condition of the contact probes (test pins).

Out of the variables analyzed, only one fully met the criteria (normality and significant part-to-part variation). Nevertheless, it is worth mentioning that the conclusions impact an equipment (DDM) which is used for a significative number of different variables, in EOL.

Additionally, significant differences in averages are generally observed across all benches, particularly between benches one and seven. Since these two do not show internal variation between nests, it is recommended to investigate the causes of drift at the bench level and perform the necessary adjustments to the measurement equipment-in this case, the DMM (e.g., performing an immediate verification and reviewing the equipment's verification frequency).

Also, it is observed that variables representing CTQ characteristics from Original Equipment Manufacturers (OEM) customers may not always be suitable for GRR studies.

Finally, the MSA study confirms the hypothesis of significant differences between ATE and confirms that reproducibility must be incorporated in future MSA evaluations.

7 Conclusions and Future Work

To build upon the findings of this study, several research avenues are proposed to further bridge the gap between measurement system variability and industrial decision-making.

A longitudinal study is proposed to correlate MSA results with quality metrics, such as internal rejection rates and customer complaints. Identifying which specific measurement variables are the strongest predictors of field failures will allow for a more targeted allocation of maintenance resources.

It is necessary to continue optimizing "MSA capable" decisions, by evaluating which variables currently measured at the EOL are truly critical for the final decision-making process, including evaluation of current CTQ variables.

If validation of the ANOVA assumption is proven through rigorous ANOVA testing, it would simplify the calibration and cross-verification protocols across the entire production line.

Future research should evaluate the impact of data quality on pass/fail decisions, by quantifying how measurement uncertainty directly impacts the probability of Type I (false rejection) and Type II (false acceptance) errors. This involves modeling the "Gray Zone" around specification limits to determine the economic and quality risks associated with the current %GRR levels.

References

1. Awad, M., Erdmann, T.P., Shanshal, Y. Barth, B.: A measurement system analysis approach for hard-to-repeat events. Qual. Eng. **21**, 3, 300–305 (2009). https://doi.org/10.1080/08982110902852344
2. Beckert, S.F., Paim, W.S.: Critical analysis of the acceptance criteria used in measurement systems evaluation. Int. J. Metrol. Qual. Eng. **8** (2017). https://doi.org/10.1051/ijmqe/2017016
3. Bordogna, L., Brembilla, F., Pagani, A., Spinetta, M.: New R&R methodology in semiconductor manufacturing electrical testing. In: Proceedings - International Test Conference, pp. 410–419 (2022)
4. Burdick, R.K., Borror, C.M., Montgomery, D.C.: A review of methods for measurement systems capability analysis. J. Qual. Technol. **35**(4), 342–354 (2003). https://doi.org/10.1080/00224065.2003.11980232
5. Fledell, E., Grossman, B.: PCB manufacturing variation impact on high frequency measurement fixtures. In: 76th ARFTG Microwave Measurement Conference: Millimeter-Wave Measurements and Modeling, ARFTG 2010 (2010)
6. Flynn, M.J., Sarkani, S., Mazzuchi, T.A.: Regression analysis of automatic measurement systems. IEEE Trans. Instrument. Measure. **58**, 10, 3373–3379 (2009). https://doi.org/10.1109/TIM.2009.2025467
7. Automotive Industry Action Group. Measurement Systems Analysis (MSA) Reference Manual. Automotive Industry Action Group (2010)
8. Guo, C., Ci, S., Zhou, Y., Yang, Y.: A survey of energy consumption measurement in embedded systems. IEEE Access **9**(2021), 60516–60530 (2021). https://doi.org/10.1109/ACCESS.2021.3074070
9. Han, Y.-J., He, Z.: An applied study of destructive measurement system analysis. In: ICIEA 2007: 2007 Second IEEE Conference on Industrial Electronics and Applications, pp. 30–34 ((2007))
10. IATF 16949. Quality management system requirements for automotive production and relevant service parts organizations. IATF (2016)
11. International Organization for Standardization (ISO). ISO/IEC Guide 98-3:2008 Uncertainty of measurement — part 3: Guide to the expression of uncertainty in measurement (GUM:1995). International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) (2008)
12. IPC-A-610G. Acceptability of Electronic Assemblies. IPC (2017)
13. ISO 5725. Accuracy (trueness and precision) of measurement methods and results — part 1: General principles and definitions. ISO (2023)

14. ISO 9001. Quality management systems — requirements. ISO (2015)
15. ISO 10012. Quality management — requirements for measurement management systems. ISO (2026)
16. ISO 26262. Road vehicles — functional safety (Series). ISO (2018)
17. ISO/IEC 17025. General requirements for the competence of testing and calibration laboratories. ISO (2017)
18. De Mast, J., Trip, A.: Gauge R&R studies for destructive measurements. J. Qual. Technol. **37**(1), 40–49 (2005). https://doi.org/10.1080/00224065.2005.11980299
19. Mazzola, S., Ara, G., Benz, T., Forsberg, B., Cucinotta, T., Benini, L.: 2025. data-driven power modeling and monitoring via hardware performance counter tracking. J. Syst. Architect. **167** (2025). https://doi.org/10.1016/j.sysarc.2025.103504
20. Mikulová, P., Plura, J., Knop, K.: Repeatibilty and reproducibility studies for non-replicable tests. Syst. Safety: Hum. – Techn. Facil. – Environ. **2020**, 275–284 (2020)
21. Raji, M., Tajary, A., Ghavami, B., Pedram, H., Zarandi, H.R.: Statistical leakage power optimization of asynchronous circuits considering process variations. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2011), pp. 126–136 (2011)
22. Razumić, A., Runje, B., Alar, V., Štrbac, B., Trzun, Z.: A review of methods for assessing the quality of measurement systems and results. Appl. Sci. (Switzerland) **15**(17) (2025). https://doi.org/10.3390/app15179393
23. Roeder, J., Altmeyer, S., Grelck, C.: Can we trust our energy measurements? A study on the Odroid-XU4. arXiv e-prints (2022). arXiv:2206.10377. https://doi.org/10.48550/arXiv.2206.10377
24. Santos, E.: Development of a method to reduce the false positive rate in the equipment's final inspection for the automotive industry. Master Dissertation. University of Minho (2023)
25. Santos, E., Silva, S., Nunes, E., Rodrigues, C.: Non-Destructive Evaluation (NDE) Of functional characteristics of electronic devices for the automotive industry: the quality of data. In: ICOPEV - 6th International Conference on Production Economics and Project Evaluation, Braga (2024)
26. Shi, L., Ren, H., Wang, J., He, Z.: Latin square design for chip length machine vision measurement system analysis. Qual. Eng. **28**(4), 381–387 (2016). https://doi.org/10.1080/08982112.2016.1147049
27. Soares, W.D.O.S., Peruchi, R.S., Silva, R.A.V., Rotella Junior, P.: Gage R&R studies in measurement system analysis: a systematic literature review. Qual. Eng. **34**(3), 382–403 (2022). https://doi.org/10.1080/08982112.2022.2069505
28. Stanković, K., Vujisić, M., Kovačević, D., Osmokrović, P.: Statistical analysis of the characteristics of some basic mass-produced passive electrical circuits used in measurements. Measure.: J. Int. Measure. Confederat. **44**(9), 1713–1722 (2011). https://doi.org/10.1016/j.measurement.2011.07.007
29. Tofaily, S., Rais, I., Anshus, O.: Quantifying the variability of power and energy consumption for IoT edge nodes. In: Proceedings - 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things, DCOSS-IoT 2023, pp. 577–584 (2023)
30. Verband der Automobilindustrie. VDA Volume 5: Measurement and Inspection Processes — Capability, Planning and Management. VDA QMC (Quality Management Center) (2021)

Radio Navigation Aids Network Optimization Using a Hexagonal Geospatial Indexing System

Ivan Ostroumov(✉) and Sergii Ivanets

Kyiv School of Economics, Mykoly Shpaka Street, 3, Kyiv 03113, Ukraine
iostroumov@kse.org.ua

Abstract. Ground-based radio navigation aid networks remain a critical component of civil aviation, ensuring the safety and reliability of air transportation. Distance Measuring Equipment (DME) is recognized as a robust backup positioning system capable of supporting the requirements of basic area navigation (RNAV) specifications. The global expansion of Free Route Airspace necessitates the optimization of DME ground networks to maintain the required levels of positioning accuracy and service continuity across all regions of the airspace. This paper proposes an algorithm for DME network enhancement based on airspace representation using a hexagonal geospatial indexing system. The algorithm identifies optimal candidate locations for ground beacon placement by maximizing the spatial distribution of positioning accuracy throughout the considered airspace. The airspace is partitioned into a set of regular hexagonal cells, enabling efficient spatial analysis. Logical operations on cell indices significantly reduce computational complexity compared to classical brute-force methods commonly used in ground network optimization. Spatial characteristic distribution of navigation system error is used to optimize network configuration. A case study of network optimization is presented using the developed software, demonstrating the effectiveness and advantages of the proposed approach.

Keywords: Air Navigation · Smart Infrastructure · Distance Measuring Equipment · Network · Intelligent Transportation System

1 Introduction

Over the past century, ground-based radio navigation aids have played a fundamental role in ensuring the safety and development of civil air transportation [1, 2]. These systems constitute a critical component of intelligent transportation system to supporting aviation operations [3, 4]. Airborne navigation equipment relies on signals transmitted by ground-based beacons to measure key parameters, such as slant range and azimuth angle. These measurements enable aircraft to follow predefined flight trajectories and to accurately determine its position in airspace.

Rapid development of Global Navigation Satellite Systems (GNSS) has led to a significant improvement in positioning performance [5, 6]. Today, GNSS serves as primary onboard positioning sensor in civil aviation due to its high levels of accuracy, availability,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 413–425, 2027.
https://doi.org/10.1007/978-3-032-30494-0_26

and continuity. In comparison, traditional ground-based radio navigation aids generally provide lower positioning performance [7]. However, vulnerability of GNSS to interference presents a critical challenge for its continued reliance in civil aviation. The number of interference events has increased dramatically over the past decade, driven by both intentional and unintentional jamming [8, 9]. Recent military conflicts, including the war in Ukraine and operations in countries of the Middle East and other regions of planet, have further underscored the importance of resilient navigation solutions and highlighted the risks associated with GNSS disruption [10]. Within the framework of Performance-Based Navigation (PBN), GNSS is considered the primary positioning source [11]. Nevertheless, in case of GNSS degradation or outage, positioning based on ground radio navigation aids or aircraft guidance using conventional systems can be employed to maintain safe flight operations. These considerations reinforce the necessity of maintaining and optimizing ground-based navigation infrastructures as a reliable backup to GNSS.

Modern roadmaps of aviation development [12] emphasize the need to maintain ground-based navigation infrastructure while promoting initiatives for its further development to enhance network configuration and positioning performance. At the same time, the air transportation system is undergoing a significant transformation toward the widespread implementation of Free Route Airspace (FRA) [13]. FRA enables aircraft to plan and operate direct trajectories between entry and exit points, particularly for en-route traffic, rather than following fixed airways. Compared to conventional airspace structures based on predefined routes, FRA offers multiple advantages, including increased operational and resource efficiency, reduced fuel consumption, and improved airspace capacity. However, its large-scale implementation requires a reconfiguration of navigational aid networks to ensure compliance with Area Navigation (RNAV) performance requirements throughout the entire airspace, rather than being limited to established flight routes [14]. Therefore, contemporary concepts of civil aviation development necessitate the modernization of navigational aid networks, including the reconfiguration of existing smart infrastructure and deployment of new nodes, to support evolving operational requirements.

Performance of navigational aid networks is typically characterized by two primary parameters: positioning accuracy and number of available station pairs [15, 16]. The number of available pairs is determined based on standard service volume model of each navigational aid. Positioning accuracy, in turn, is evaluated using standardized methodologies defined by ICAO, ensuring a globally consistent assessment framework applicable to aircraft operating in any region. These performance indicators are then compared against the minimum requirements specified by RNAV requirements. However, no analytical method exists for directly estimating spatial distribution of these parameters across a continuous airspace. As a result, existing approaches rely on discretizing airspace into a set of cells, followed by evaluation of performance metrics for each individual cell. Most current methods adopt brute-force strategies for parameter estimation within each cell [17, 18]. The primary distinction between these methods lies in the strategy used to traverse or prioritize the cells during analysis phase. Techniques such as sequential evaluation, gradient-based methods, and other variations of exhaustive search are commonly applied in airspace performance analysis.

The problem of optimal navigational aid network configuration is typically addressed through brute-force selection of candidate locations, followed by the maximization of an objective function aimed at achieving required positioning accuracy across the largest possible portion of airspace. However, such approaches require repeated evaluation of network performance metrics for every discretized cell within the considered airspace, resulting in high computational demands and the need for substantial processing resources.

Over the past decades, global indexing systems have undergone significant development, evolving from simple visualization libraries to sophisticated frameworks that provide advanced tools for navigation and geospatial data aggregation. The hierarchical structure of addressing systems facilitates efficient data aggregation, enabling adaptation to required level of spatial precision and available computational resources. Approaches such as Open Location Codes and Hierarchical Hexagonal Indexing [19, 20] demonstrate substantial potential for geospatial analysis. These global indexing systems are widely applied in data analysis tasks where parameters are spatially distributed across geographic areas [21, 22].

In this paper, we consider an approach for effective network configuration estimation based on available areas of potential nodes placement and local impact study in network configuration manipulation. Proposed approach incorporates both positioning accuracy and its spatial distribution as key parameters within the optimization objective function. The airspace is discretized using a hexagonal spatial indexing system, which significantly reduces computational complexity at each iteration step. This representation also enables efficient local adjustments of the network configuration, facilitating the identification of optimal locations for ground beacon deployment.

2 Network of Radio Navigation Aids

There are two basic types of navigational aids used in civil aviation: distance measuring equipment (DME, also known as Tactical Air Navigation or TACAN) and VHF omnidirectional ranging (VOR) [23]. DME measures range between the airplane and each node of DME ground network. VOR is measuring an azimuthal angle to ground station from airplane position. There are also non-directional beacons (NDB) ground networks, but their configuration is reduced significantly due to low precision of angle measuring equipment on board of airplane.

DME is an active measuring device. DME interrogator is an avionics part which generates interrogation signals and receives replies from ground stations. The interrogator measures time traveling in space and calculates range between airplane and ground beacon. Each ground beacon operates at a specific radio channel (there are 256 DME channels specified in normative documents). All channels are frequency distributed. Each DME ground station receives integration signals and, after a fixed time of delay, generates a reply at another co-located radio frequency. Ground DME network forms a graph with exactly known location of each node.

VOR has another operation principle. VOR ground station is continuously transmitting a navigation signal that includes two main components. One component has information about reference direction (North), and another component includes information about direction. Therefore, avionics part of VOR has only a receiver, which

includes amplitude and frequency demodulators. Both demodulated signals are compared to identify the shift of phases between them, which is proportional to measured angle.

Provider of air navigation services in each country is responsible in navigational aids network deployment and maintaining specified network parameters. Both DME and VOR networks are usually considered separately, however VOR/DME navigation could also be provided based on co-allocated VOR/DME beacons. An example of navigational aids networks is given in Fig. 1.

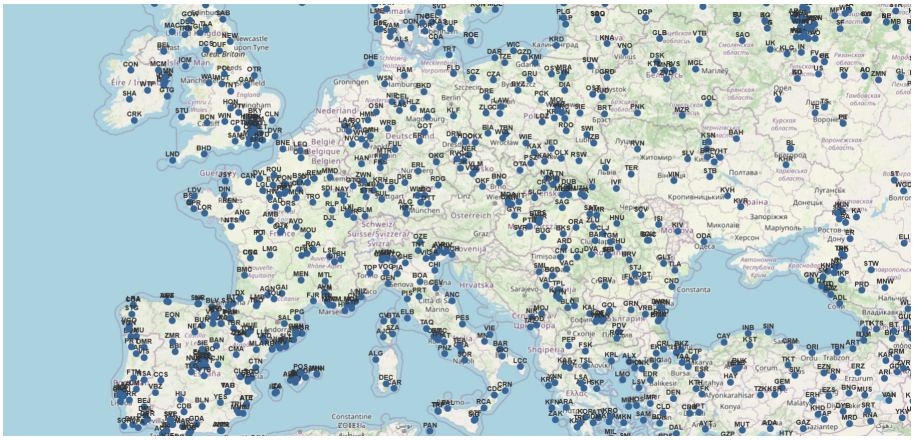


Fig. 1. Nodes of DME and VOR networks in Europe.

Low performance of azimuthal angle measuring in VOR significantly limits performance of VOR/VOR positioning method. Results of the previous study show that VOR/VOR ground facility could not support navigation by level of RNAV 1, which is required for continental flights [24, 25]. In contrast, DME/DME navigation fully supports RNAV 1 precision level and could guarantee safe air navigation [26, 27]. It makes DME/DME navigation useful for further consideration instead of VOR/VOR and makes DME network perspective, and VOR marked as a back-up capability which should be minimized to guarantee only the basic coverage in case of emergency.

In the common case, the network of navigational aids is represented as a graph with beacons placed in nodes. Beacons for which service areas are crossing an edge makes a connection between nodes.

3 Algorithm of Network Efficiency Improvement

This study considers the scenario in which an existing navigational aids network is already deployed, and objective is to identify optimal locations for installation of additional ground beacons. This problem is particularly relevant for air navigation service providers seeking to enhance network performance in accordance with modern development roadmaps. As input, the proposed approach assumes a predefined set of feasible

regions for ground station placement, reflecting practical constraints such as terrain, smart infrastructure availability, and regulatory limitations.

The airspace considered for network optimization is typically defined by administrative boundaries of a country or by limits of a Flight Information Region (FIR). The configuration of FIRs is established in regulatory documents and maintained through international agreements between states. In practical implementations, it is often more efficient to utilize standardized ICAO FIR identifiers to retrieve the corresponding airspace boundaries required for analysis.

Within the administrative boundaries of a FIR, the airspace is partitioned into hexagonal cells. The global hierarchical hexagonal indexing system (H3) provides an efficient and flexible framework for such spatial representation. H3 supports 17 hierarchical levels of precision, each defined by hexagonal cells of a specific size. The highest precision level corresponds to cells with an area of approximately 0.9 m^2[28]. Each H3 cell is assigned a unique address, which also encodes addresses of all its child sub-cells. The regular hexagonal geometry of H3 cells facilitates efficient computation of spatial properties, including area calculation, distance measurements, and the identification of all cells within a given radius, making it particularly suitable for airspace analysis and network optimization tasks.

H3 library provides flexible geo-coding by input polygon of coordinates. For network analysis, each hexagonal cell is assumed to have uniform network parameters, and the coordinates of the cell center are used as representative points for evaluating performance metrics. This approach simplifies computation while maintaining sufficient accuracy for network optimization.

Each network node is associated with a specific H3 cell index. There are two primary approaches for evaluating service volume of a navigational aids. The first approach is based on radio wave propagation equations and the maximum achievable range of communication link. The second approach employs standardized service volumes, typically approximated using cylindrical models. Often, multiple cylinders are combined to represent the service volume over varying terrain. Simplified cylindrical models are widely used in air navigation due to its computational efficiency, particularly for onboard equipment. Accordingly, adopting cylindrical models is a practical and effective choice for identifying candidate locations for new beacon placement.

Normative documents describe only four DME beacon types which could serve en-route phase of flight, low altitude (with 40NM radius), high altitude (130 NM), DME Low (130 NM) and DME High (130 NM) [29]. Each of these types has a different combination of cylinders with exponential approximation at low altitude.

In a computer simulation, a fixed altitude could be used for identification of radius of involved area around beacon. Functionality of H3 library includes a k-ring algorithm, which identifies indexes of all cells placed in the disk with a specified center and radius. It requests that all network nodes should be associated with cell indexes and radius of service area at considered altitude layer, transformed to the metric of parameter k.

The number of available beacons for each cell is estimated as intersection of sets as follows:

$N_i = \bigcap_{j=1}^{n} DME_j^i$ (1).

where n is number of cells; DME^i is a binary set of availability for i-th beacon.

A set of possible combinations of two is estimated for each available number of beacons at a particular cell. Each pair of beacons in combination is considered separately to estimate Total System Error (TSE). A pair with the lowest TSE is used as the most productive.

In case of DME/DME navigation TSE is estimated as follows:

$TSE^2 = FTE^2 + 4\frac{2\sigma_s^2+\sigma_{int1}^2+\sigma_{int2}^2}{sin(\alpha)^2}$ (2).

where FTE is flight technical error; α is inclusion angle at point of airplane location; σ_s is standard deviation error caused during communication line and uncertainty of beacons delays; σ_{int} is the standard deviation error of avionics interrogator of DME.

FTE is caused by an error of piloting mode used for airplane guidance. For the case of autopilot mode, which is most common for en-route phase of flight FTE = 0.25 NM.

Maximum value of standard deviation error caused by communication line and uncertainty of beacon delays is specified in normative documents at a level of 0.05 NM [11].

Standard deviation error of DME interrogator is distance-related and usually used as follows [11]:

$\sigma_{int} = max(0.125D, 0.085NM)$ (3).

where D is measured range between airplane and DME beacon.

Estimated TSE value at each cell is compared to peril value of 1 NM. If accuracy limit is reached, then binary flag is issued to a particular cell. Trigger function is as simple as following:

If $TSE_i < 1$, then $RNAV_i = 1$, else $RNAV_i = 0$. (4).

Algorithm uses each cell from available pull of possible new beacon locations as a candidate for network node placement with a particular DME type. Algorithm required recalculation of Eqs. (1–3) only for cells placed in disc cells centered at the candidate location. It does not require recalculation of whole network but only cells within the availability disc. Results of TSE triggering function (4) identify the number of additional cells that could be obtained from application candidate location. Thus, number of additional RNAV 1 valid cells is a good metric to identify the best candidate for node placement.

The search process considers each potential cell and estimates its impact in the overall network, based on number of additional RNAV 1 complement cells. Finally, the decision is given to place, which provides the maximum number of valid cells.

The structural scheme of proposed algorithm is given in Fig. 2. Initiation of search requires a connection to FIR boundaries database and navigational aids database. Also, setting up a precision level and specification of areas of possible beacon locations should be given in the beginning. Based on software realization scheme, areas of possible beacon locations could be specified by an interactive user interface and a map projection library.

The task of existing network reconfiguration could also be solved by proposed method by sequentially deleting one of the nodes and considering the network of low functionality with a search for the best place for additional beacons. Thus, proposed approach could be used not only for adding a new node, but also for updating the existing network configuration.

For multiple beacons, there are two strategies that could be applied: the first one is based on sequentially adding each new node to the network. The second one considers

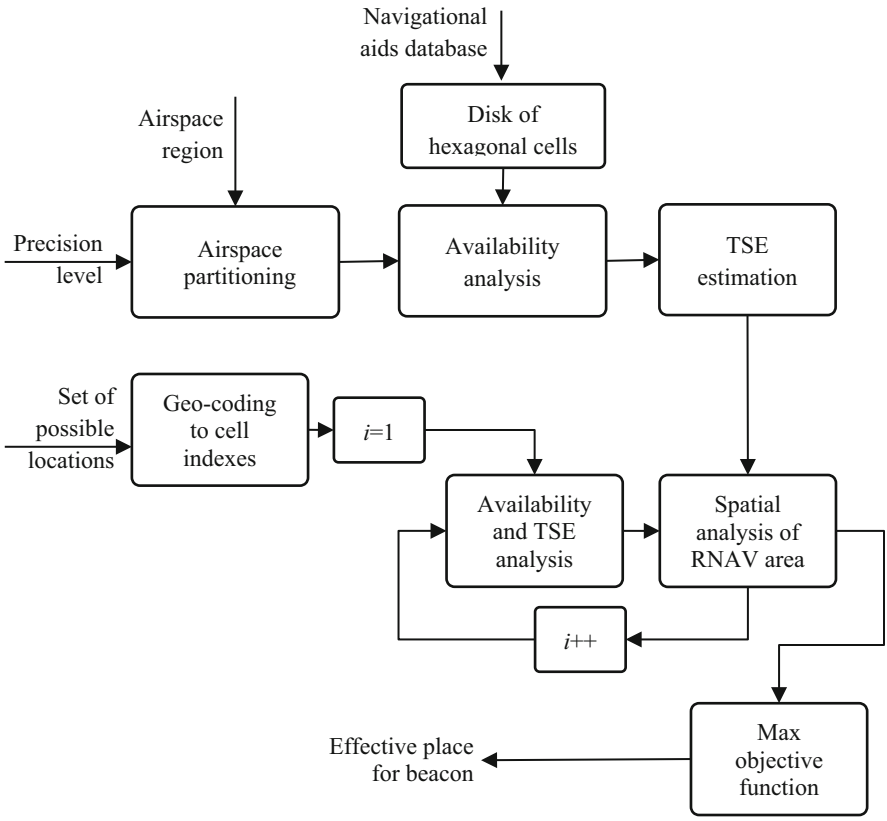


Fig. 2. Structural scheme of proposed algorithm for effective network improvement.

scenes with all possible combinations of beacon injection into the network. The last one is the most precise, but is the most computationally heavy. Both approaches are based on applying the same proposed algorithm, but work with a different set of cell indexes placed on the disks.

It makes sense to run the algorithm several times at a low level of precision first and then with improved precision for results identified in the rough first scanning phase. It works best if potential places for beacons are not specified and the whole area of airspace is considered as possible places for node placement.

4 Navigation Aids Network Improvement

In the numerical case study, the proposed method is applied to optimization of the navigational aids network of Ukraine with the objective of improving overall service coverage. War in Ukraine caused significant impact into configuration of navigational aids network [30, 31]. Due to reduced network functionality in the eastern and southern regions of considered airspace, the analysis focuses on identifying optimal locations for additional beacons in the central part of the airspace, where network enhancements

can provide the greatest overall performance benefit. Ukrainian airspace consists of five FIRs: UKLV, UKBV, UKOV, UKDV, and UKFV. Also, in this task, we consider only DME network with 16 ground beacons placed at different FIRs [32]:

- UKLV: UZH (Uzhhorod), IVF (Ivano-Frankivsk), LIV (Lviv), TER (Ternopil), RVN (Rivne);
- UKBV: VIN (Vinnytsia), SLV (Soloviivka), STB (Stebliv), BRP (Boryspil), KSN (Koshany), YHT (Yahotyn), BAH (Bakhmach), KVH (Kropy);
- UKOV: ODA (Odesa);
- UKDV: KHR (Kharkiv), KVR (Kryvyi Rih).

But this network is a part of neighboring airspace networks: Poland (42 DMEs), Slovak Republic (12 DMEs), Hungary (9 VOR/DMEs), Romania (24 DMEs), Moldova (1 VOR/DMEs), Bulgaria (12 DMEs), and Turkish (60 DMEs).

For practical realization, a software has been developed in JavaScript to validate the proposed algorithm [33]. Software for a browser is a good example of a cross-platform application. In addition, a variety of mapping libraries are available to support easy visualization of geographic data in different map projections. Also, H3 library has a friendly interface for integration in JavaScript solutions. A simple interface has been created to make access to set up network configuration and interact with geographic data. Basic features of software have been developed in PHP with MySQL to host navigational aids database and validate the security reasons for data protection from unauthorized access. All navigation calculations and algorithms are realized in user interface. Modern hardware has made possible to run big data analysis on common equipment. In case of low productivity, the lowest precision level should be used to make time of computation visible.

Results of network availability estimation are given in Fig. 3 for the initial network configuration. Total area of considered region is 790×10^3 km^2. For simulation, we use 7^{th} level of H3 precision with an effective area of cell in 36 km^2. Services of 73 beacons are available in considered airspace. Area of DME service is 740.6×10^3 km^2 (which is above 93% of whole considered airspace).

Results of navigation system error estimation are given in Fig. 4 for initial network performance. In simulation, an area of possible candidates for ground facility installation is given as polygon shape. Restrictions on places for installation are important, because it does not make any sense to run optimization and get places where a ground station cannot possibly be deployed due to physical (terrain or water surface) or legal (restrictions due to private property) reasons. Therefore, specifying a polygonal area (or many areas) is important to get designed results. In our simulation, we consider a scenario of best candidate search for a node in UKDV FIR. And possible locations have been proposed in the vicinity of Dnipro city. Also, considering new places for beacons deployment is important to consider interference environment and the given logistical and technical infrastructure (for example, available power supply and availability of wired communication lines for stable remote connection).

Results of analysis show that in current network configuration, a significant improvement of performance from adding only one additional node could be obtained in the vicinity of Dnipro airport. Identified cell address is “862d64da7ffffff” (with a center point at 48.3605203N, 35.1102016E), which is highlighted by a red circle in Fig. 5.

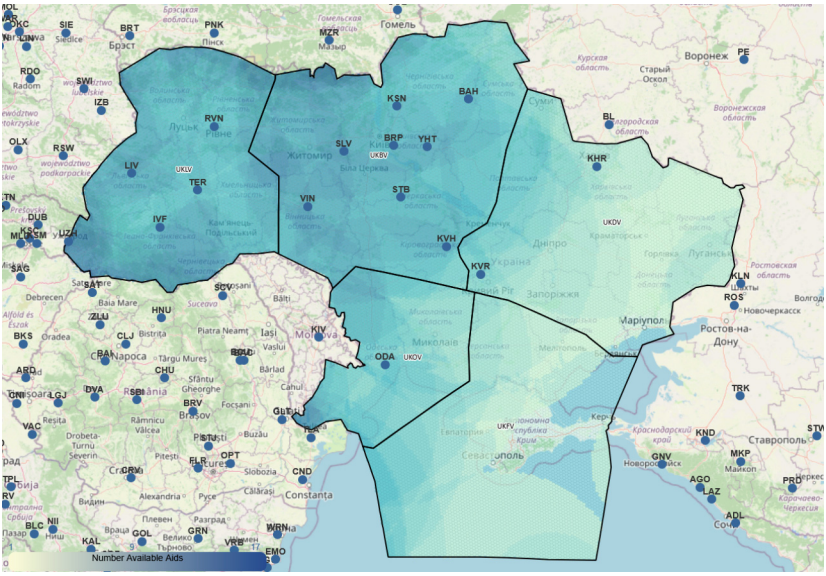


Fig. 3. Results of availability estimation for initial DME network configuration.

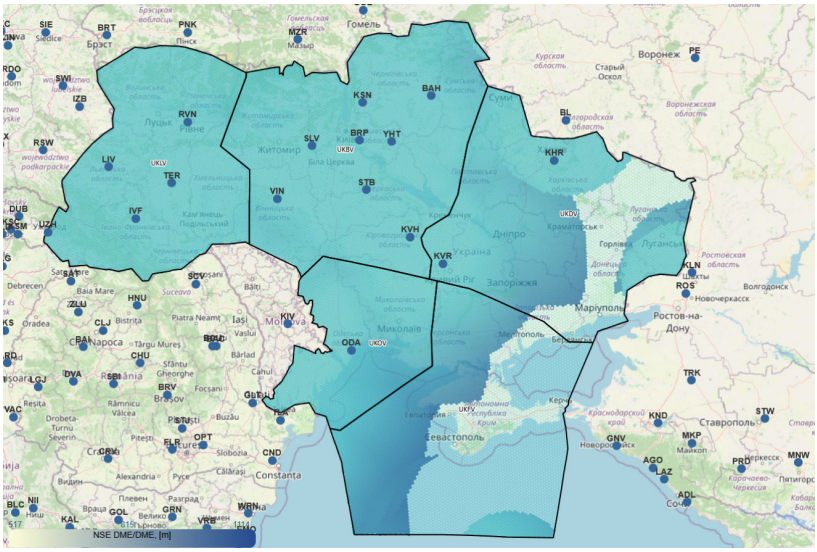


Fig. 4. Initial DME network performance.

Results of estimation of spatial characteristics for current and improved network performance are given in Fig. 6 for availability, and Fig. 7 presents results in navigation system error (NSE).

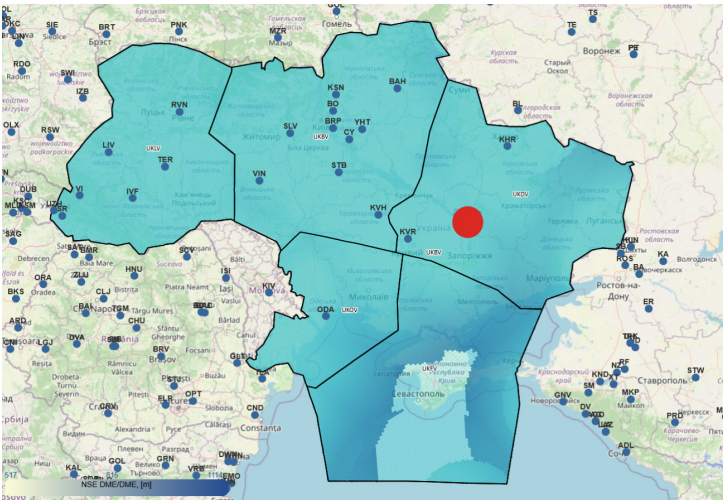


Fig. 5. Identified location of DME.

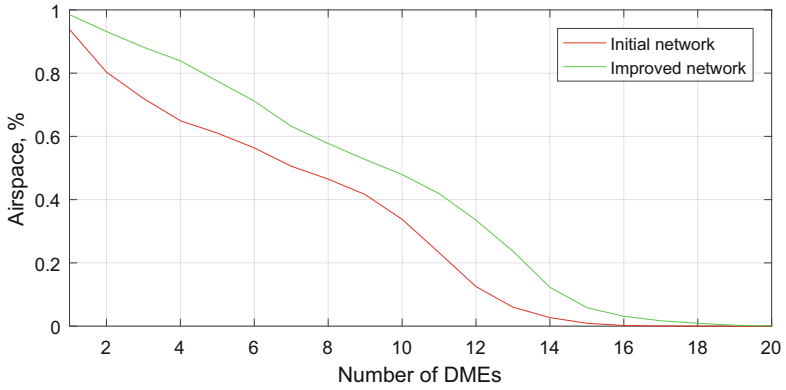


Fig. 6. Results of improved network availability.

Obtained results indicate that placement of only one additional node to the network configuration could result in significant network improvement. The best results could be reached in the peril zones of network when additional DME could be fruitful in the pair formation process. Obtained results show significant improvement in spatial distributions of parameter availability and performance of network service.

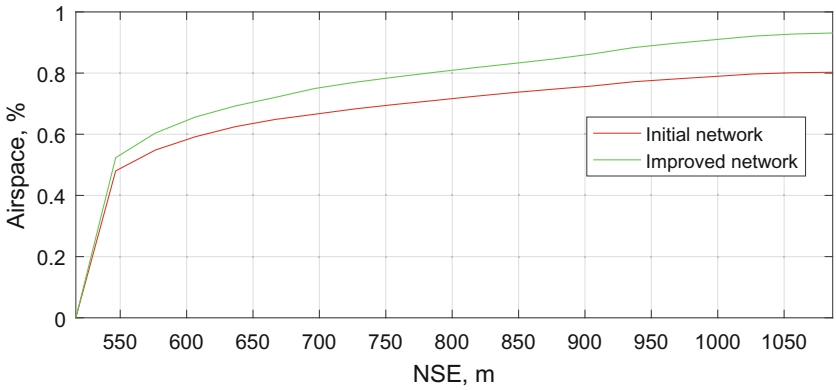


Fig. 7. Results of improved network positioning performance in NSE.

5 Conclusions

Network optimization is one of the core tasks of modern air transportation development in order to improve the safety of civil aviation. Improvement of positioning performance by navigational aids could be reached by adding new nodes in the existing network. H3 library provides a new step in tools for geo-data analysis with easy access to cell properties calculation. In comparison with existing methods, proposed algorithm uses a reduced number of recalculation at each candidate point analysis, which requires recalculation only effected cell with no need for recalculation of the whole grid of cells. In addition, different precision levels of hexagons could be used for rough scanning and precise tuning results.

Results of network improvement depend on quality of existing network configuration. Nodes type and the type of each addition navigation aid have a direct impact on the overall optimization process. Proposed algorithm provides resource efficiency distribution within navigation aids network to support modern trends in civil aviation development.

Practical application of the proposed method for Ukrainian navigational aids improvement shows that one additional node could significantly improve overall network performance for considered polygon of airspace.

References

1. Lo, S., Enge, P., Niles, F., Loh, R., Eldredge, L., Narins, M.: Preliminary assessment of alternative navigation means for civil aviation. In Proceedings of the 2010 International Technical Meeting of The Institute of Navigation, pp. 314–322 (2010)
2. Espen, M., Schafer, M., Vitan, V., Berz, G., Eichhorn, R., Schmitt, J.: Old but gold: evaluating the accuracy and integrity of DME based on real-world measurements. In: 2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC), pp. 1–10. IEEE, USA (2024)
3. Ostroumov, I., Kuzmenko, N., Zaliskyi, M.: Global assessment of radio navigation aid networks and their contribution to performance-based navigation implementation. Eng **12**(6), 360 (2025). https://doi.org/10.3390/eng6120360

4. Espen, M., Schafer, M., Vitan, V., Berz, G., Eichhorn, R., Schmitt, J.: Old but still gold: identifying and investigating multipath propagation effects in DME. In 2025 AIAA DATC/IEEE 44th Digital Avionics Systems Conference (DASC), pp. 1–10. IEEE, USA (2025)
5. Maciejewska, A., Maciuk, K.: Research using GNSS (Global Navigation Satellite System) products–a comprehensive literature review. J. Appl. Geodesy **19**(4), 555–574 (2025)
6. Joubert, N., Reid, T. G., Noble, F.: Developments in modern GNSS and its impact on autonomous vehicle architectures. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 2029–2036. IEEE (2020)
7. Ostroumov, I.V., Kuzmenko, N.S.: Risk analysis of positioning by navigational aids. In: 2019 Signal Processing Symposium (SPSympo), pp. 92–95. Krakow, Poland (2019). https://doi.org/10.1109/SPS.2019.8882003
8. Felux, M., Fol, P., Figuet, B., Waltert, M., Olive, X.: Impacts of global navigation satellite system jamming on aviation. Navigation **71**(3), 1–22 (2024). https://doi.org/10.33012/navi.657
9. Li, X., et al.: Overview of jamming technology for satellite navigation. Machines **11**(7), 768 (2023)
10. Figuet, B., Waltert, M., Felux, M., Olive, X.: GNSS jamming and its effect on air traffic in eastern Europe. Eng. Proc. **28**(1), 12 (2022)
11. ICAO. Performance-Based Navigation (PBN) Manual. Doc 9613, 5th ed.; ICAO: Montreal, QC, Canada (2023)
12. European ATM Master Plan. EUROCONTROL. https://www.eurocontrol.int/portal/european-atm-master-plan-portal. Accessed 10 March 2026
13. Brown, M., et al.: Introducing free route airspace to northeast asia/pacific region. Trans. Japan Soc. Aeron. Space Sci. **67**(6), 350–361 (2024)
14. Lopez-Lago, M., Serna, J., Casado, R., Bermudez, A.: Present and future of air navigation: PBN operations and supporting technologies. Int. J. Aeronaut. Space Sci. **21**(2), 451–468 (2020)
15. Almogbil, T.S., Yahyia, A.A., Alturki, A.M., Qttlan, S.A.: Enhancing PBN coverage with DME optimization and development. In: Integrated Communications. Navigation and Surveillance Conference (ICNS), pp. 1–4. IEEE, Brussels (2025)
16. Lo, S., Chen, Y.H., Enge, P., Peterson, B., Erikson, R., Lilley, R.: Distance Measuring Equipment Accuracy Performance Today and for Future Alternative Position Navigation and Timing (APNT), pp. 711–721. Nashville, TN (2013)
17. Topkova, T., Pleninger, S., Hospodka, J., Kraus, J.: Evaluation of DME network capability using combination of rule-based model and gradient boosting regression. J. Air Transp. Manag. **119**, 102637 (2024)
18. Kim, E.: Investigation of APNT optimized DME/DME network using current state-of-the-art DMEs: Ground station network, accuracy, and capacity. In: Proceedings of the 2012 IEEE/ION Position, Location and Navigation Symposium, pp. 146–157. Myrtle Beach, SC, USA (2012). https://doi.org/10.1109/PLANS.2012.6236876
19. Uher, V., Gajdos, P., Snasel, V., Lai, Y., Radecky, M.: Hierarchical hexagonal clustering and indexing. Symmetry **11**(6), 731 (2019). https://doi.org/10.3390/sym11060731
20. Amiri, A.M., Samavati, F., Peterson, P.: Categorization and conversions for indexing methods of discrete global grid systems. ISPRS Int. J. Geo Inf. **4**(1), 320–336 (2015). https://doi.org/10.3390/ijgi4010320
21. Morris, J., Lee, S.M., Chinthavali, S., Bhusal, N., Ahmed, N., Iyer, G.: HexWeather: hexagonal spatial data aggregation for weather-driven grid resilience analysis. In International Conference on Information Reuse and Integration and Data Science (IRI), pp. 67–72. IEEE (2025)

22. Wojtkiewicz, K., Palak, R., Jodlowiec, M., Hasiak, M., Mikłasz, J.: Open hexagon-based air pollution prediction service with public API and embeddable web widget. In: Nguyen, N.T., et al. Computational Intelligence in Engineering Science. ICCIES 2026. Communications in Computer and Information Science, vol. 2944. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-21628-1_42
23. Aeronautical Telecomunications. Radio navigation aids. Annex 10 to the Convention on International Civil Aviation, vol. I, 8th edn. ICAO, Montreal, Canada (2023)
24. Ostroumov, I.V., Kuzmenko, N.S.: Compatibility analysis of multi signal processing in APNT with current navigation infrastructure. Telecommun. Radio Eng. **77**(3), 211–223 (2018). https://doi.org/10.1615/TelecomRadEng.v77.i3.30
25. Ostroumov, I.V., Kuzmenko, N.S.: Accuracy improvement of VOR/VOR navigation with angle extrapolation by linear regression. Telecommun. Radio Eng. **78**(15), 1399–1412 (2019). https://doi.org/10.1615/TelecomRadEng.v78.i15.90
26. Harbeke, J., Espen, M., Schafer, M., Vitan, V., Berz, G., Schmitt, J.: A data-driven comparison of DME-based navigation performance. In: 44th Digital Avionics Systems Conference (DASC), pp. 1–10. AIAA DATC/IEEE, Montreal, QC, Canada (2025). https://doi.org/10.1109/DASC66011.2025.11257319
27. Strecha, P., Makula, P.: Comparison of simple DME/DME positioning method and augmentation using extended kalman filtering, unscented kalman filtering and particle filtering. In: 38th Digital Avionics Systems Conference (DASC), pp. 1–6. IEEE/AIAA (2019)
28. Hexagonal Hierarchical Geospatial Indexing System Specification. https://h3geo.org. Accessed 10 March 2026
29. U.S. Terminal and En Route Area Navigation (RNAV) Operations, Advisory Circular no. 90-100A. U.S. Department of Transportation. FAA (2014)
30. Ostroumov, I., Ivashchuk, O., Kuzmenko, N.: Preliminary estimation of war impact in ukraine on the global air transportation. In: 12th International Conference on Advanced Computer Information Technologies (ACIT), pp. 281–284. IEEE, Ruzomberok, Slovakia (2022). https://doi.org/10.1109/ACIT54803.2022.9913092
31. Brik, T., Mylovanov, T., Murtazashvili, J.B., Murtazashvili, I.: Introduction: special issue on the political economy of the War in Ukraine. J. Public Finan. Public Choice **39**(1), 2–9 (2024). https://doi.org/10.1332/25156918Y2024D000000006
32. AIP Ukraine, 2024. Aeronautic Information Publication of Ukraine. UkSATSE
33. Airspace Analysis. https://network.sciary.com. Accessed 10 March 2026

Hybrid Multi-logic Reasoning Framework Over Ontologies for Decision Support in Neurodevelopmental Disorders

William Chaves de Souza Carvalho[1(✉)], Pedro Moisés de Souza[2], and Flávio de Oliveira Silva[3]

[1] Federal University of Uberlândia, Campus Santa Mônica, Uberlândia, Brazil
william.carvalho@ufu.br
[2] Federal University of Viçosa, Campus Rio Paranaíba, Rio Paranaíba, Brazil
pedromoises@ufv.br
[3] Universidade do Minho, Braga, Portugal
flavio@di.uminho.pt

Abstract. Neurodevelopmental disorders pose significant challenges for clinical decision support because their assessment involves heterogeneous data, evolving developmental trajectories, uncertainty, and, at times, conflicting evidence. In this paper, we present the Hybrid Multi-Logic Autism Ontology framework (HML-AutismOnt), a modular, layered architecture developed to extend the Autism Ontology. We designed this extension to address the limitations of traditional decision support systems in Neurodevelopmental Disorders (NDDs) management, particularly the difficulty of representing temporal evolution, conflicting clinical evidence, and normative constraints within care pathways. The framework integrates temporal, fuzzy, modal, deontic, and paraconsistent logics to model evolving states, uncertain observations, guideline-driven obligations, and potentially inconsistent evidence. We describe the system architecture, the hybrid reasoning workflow, and the organizational strategy that combine ontology-based knowledge representation with modular reasoning services. The resulting framework promotes semantic interoperability, enables structured multi-logic inference, and supports consistent reasoning over complex domain knowledge. More broadly, this work contributes to the design of semantic information systems capable of handling heterogeneous, uncertain, and temporally evolving data in real-world neurodevelopmental care.

Keywords: Ontology Engineering · ontology enrichment · Hybrid Reasoning · Knowledge Representation · Decision Support Systems · Semantic Information Systems · Neurodevelopmental Disorders · Non-classical Logics

1 Introduction

Neurodevelopmental Disorder (NDDs) constitute a heterogeneous group of conditions arising from atypical brain development and significantly affect personal,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 426–445, 2027.
https://doi.org/10.1007/978-3-032-30494-0_27

social, academic, and occupational functioning [8] and [25]. Among them, Autism Spectrum Disorder (ASD) stands out for its high prevalence, complex clinical presentation, and substantial impact on individuals, families, and healthcare systems [8,22].

Specifically, according to American Psychiatric Association (APA) and World Health Organization (WHO), ASD is a neurodevelopmental condition characterized by persistent early-onset deficits in social communication and restricted or repetitive behaviors. As a spectrum disorder, it involves graded difficulties in these domains, which can result in functional limitations in everyday life. Its severity is classified into three levels of support (1 to 3), indicating the degree of assistance required [8,31].

The increasing prevalence of autism, compared with other NDDs, has not only widened the discrepancy in the number of cases, but also amplified educational, clinical, and computational challenges. This mounting pressure highlights the urgent need for semantic models that can: *i*) integrate diverse clinical and contextual data; *ii*) capture variability, uncertainty, and potentially conflicting evidence in diagnosis; and *iii*) support individualized adjustments as described in the Diagnostic and Statistical Manual of Mental Disorders, 5th Ed., Text Revision (DSM-5-TR) [8,9,29].

Real-world challenges, for instance, pose greater computational challenges. NDDs pose significant challenges from a computational perspective, particularly due to their intrinsic heterogeneity, temporal evolution, and the frequent coexistence of uncertain or even conflicting clinical evidence. These characteristics make it difficult for traditional ontology-based decision support systems, especially those grounded in classical Description Logic (DL), to adequately represent and reason in such domains. Within this broader context, ASD emerges as a particularly demanding case, given its high prevalence, complex clinical presentation, and substantial impact on individuals, families, and healthcare systems [22].

We are long ago convinced that overcoming such limitations requires robust semantic technologies. Ontologies have long been a powerful computational framework for the biomedical field. The Unified Medical Language System (UMLS), in 1993, laid the groundwork for semantic integration across heterogeneous biomedical sources [21], while in 1996 the GALEN ontology was among the first to apply formal ontological principles specifically to the representation of clinical and biomedical knowledge [28]. In 2015 and 2016, the first two works that formally use ontologies, in the computational/logical sense, using OWL or similar, specifically for neurodevelopmental disorders, appear [16,23]. Although formal ontologies have been developed for other NDDs, such as Attention-Deficit/Hyperactivity Disorder (ADHD) [6], specific learning disorders, including dyslexia [27], and intellectual disability [19], their application to the autism domain remains particularly challenging.

However, ontologies based on DL, thus classical logic, face significant limitations when applied to ASD. Symptom severity is inherently graded rather than binary [3,24]; developmental trajectories require explicit temporal model-

ing [15,20]; clinical decisions often involve normative and ethical considerations [26,30]; differential diagnoses rely on notions of possibility and necessity [12,34]; and multi-source assessments frequently contain genuine contradictions [10,17]. These characteristics exceed the expressive capacity of standard Web Ontology Language (OWL) 2 DL and Semantic Web Rule Language (SWRL), thus limiting the precision, adaptability, and clinical utility of existing ontologies. However, although new frontiers in representation can be reached by integrating non-classical logic into ontologies, compatibility with OWL 2 DL must be preserved for decidability [7].

To address these limitations, we focused specifically on ASD and introduced Hybrid Multi-Logic AutismOnt (HML-AutismOnt), a hybrid multi-logic framework that systematically extends the AutismOnt ontology through a modular layered architecture [14]. AutismOnt is a comprehensive ontology for ASD that supports automated inference via SWRL rules and has been applied to clinical decision support, text annotation, and interdisciplinary research [5,14,32]. By integrating non-classical logics, including fuzzy, paraconsistent, temporal, modal, and deontic reasoning, the proposed architecture significantly enhances the semantic expressivity of the ASD characteristics. The resulting models are more nuanced and clinically actionable, overcoming the inherent limitations of traditional first-order ontological approaches while preserving full interoperability with the original AutismOnt.

This paper makes the following contributions: *i*) the extension of AutismOnt with a hybrid multi-logic representation framework that captures heterogeneous and evolving domain knowledge; *ii*) the integration of five complementary non-classical logics—modal, temporal, fuzzy, deontic, and paraconsistent—within a unified reasoning model; *iii*) the definition of a hybrid rule-based mechanism Hybrid Semantic Web Rule Language (H-SWRL) that enables cross-logic inference over annotated ontology assertions; *iv*) the design of a modular semantic architecture that supports the coordinated execution of heterogeneous reasoning services; and *v*) Demonstration of the framework's applicability through a longitudinal decision-support scenario involving complex, evolving data.

The remainder of this article is structured as follows. Section 2 synthesizes the relevant previous work. In Sect. 3, we present the proposed framework. Section 4 presents a detailed illustrative hypothetical example through a longitudinal case study of a ninth-year-old boy's assessment. Section 5 examines the implications of the findings and delineates the limitations of the study. Finally, Sect. 6 offers concluding remarks.

2 Related Work

In this section, we examine ontology-based methodologies that have become pivotal for the systematic organization and semantic integration of knowledge in the field of ASD and other NDDs. In recent years, several influential ontology-driven frameworks and models have been proposed. We highlight their principal achievements and clarify how the architecture we developed both builds

upon and advances this body of work by concurrently incorporating five non-classical logical formalisms. To situate our proposal within the existing research landscape, we begin by discussing foundational approaches that are primarily formulated within the classical logic paradigm.

A seminal contribution in this area is the Autism DSM-ADI-R Ontology (ADAR) developed by [23], which was among the first systems to employ OWL and SWRL rules to infer ASD phenotypes directly from Autism Diagnostic Interview-Revised (ADI-R) assessment data. This ontology demonstrated high precision in aligning patient responses with the diagnostic criteria of DSM-5-TR. The study provided compelling evidence for the practical utility of formal ontologies in phenotype extraction. However, the approach was limited to the logic of classical DL, which offers only limited capabilities for representing the severity of symptom classification and for handling inconsistent or conflicting clinical information.

AutismOnt was introduced by [14] and remains the most comprehensive ontology exclusively dedicated to ASD. Developed in Protégé using the OWL 2 DL formalism, it comprises 676 classes and more than 124 properties, covering diagnostic criteria, phenotypic manifestations, risk factors, interventions, comorbid conditions, and aspects of life. The ontology supports automated reasoning through SWRL rules and has been successfully applied to clinical decision support systems and text annotation tasks. Although it provides a comprehensive, structured foundation, its dependence on crisp bivalent logic limits its ability to model the inherent vagueness, temporal dynamics, and contradictory information frequently encountered in ASD assessments. We therefore adopted AutismOnt as the foundational ontology of our framework and extended it while preserving full backward compatibility.

In our survey of recent partial hybrid methodologies, we observed that the work of [4] introduced a composite framework that integrates ontological representations with machine learning algorithms and multi-criteria decision-making techniques to enhance the diagnostic precision of ASD, explicitly incorporating sociodemographic variables into the analysis. Their approach exhibited promising results in integrating heterogeneous data sources, but the hybridization remains partial and does not incorporate non-classical logics to handle graded severity or ethical constraints. Another notable work on partial hybrid approaches is Autism Spectrum Disorder Knowledge Base (AsdKB) created by [33], a Chinese knowledge base that focuses on screening for ASD and providing expert recommendations. The system integrates ontological knowledge with factual data in Resource Description Framework (RDF) format, providing valuable multilingual support. However, it also operates primarily within classical logic, leaving room for extensions that address temporal evolution and contradictory evidence.

These partial hybrid approaches have led to recent works that incorporate specific non-classical elements. For example, [18] presented a hybrid reasoning model for the care of ASD that integrates rule-based, case-based, and process-mining reasoning within an ontological framework. Their system is particularly

useful for predicting behavioral trajectories and supporting intervention planning. Nevertheless, it does not incorporate fuzzy, modal, or deontic logics, which limits its capacity to model the severity, possibility, and normative aspects of symptoms simultaneously. Recently, [4] developed an intelligent fuzzy-neural framework specifically for sensory assessment in autism, using hierarchical linguistic modeling and risk-based temporal decision-making. [4] takes an important step toward tackling vagueness in sensory processing, but its scope is restricted to fuzzy and temporal aspects, leaving paraconsistent, modal, and deontic reasoning entirely unexplored.

Within the broader framework of hybrid logics applied to biomedical domains, composite systems integrating multiple non-classical logics have been investigated across various medical specialties, producing promising and empirically encouraging outcomes. Fuzzy-temporal approaches have been used to model disease progression in cardiology [11], while paraconsistent logics have helped to manage contradictory laboratory findings in oncology [1]. Modal and deontic logics have been combined to support the compliance of ethical guidelines in intensive care units [13]. Despite these advances in adjacent domains, no existing work in the investigation of NDDs or ASD has achieved the full concomitant integration of modal, temporal, fuzzy, deontic, and paraconsistent logics within a single ontology. The HML-AutismOnt framework fills this gap, occupying a prominent position at the cutting edge of both scientific interest and computational innovation, standing at the forefront of advanced research in the field. Table 1 summarizes the main related studies.

We have structured Table 1 so that each column corresponds to a parameter of interest related to the analyzed work. Regarding the table, we will use the terms *parameter* and *column* interchangeably. Thus, the scope of the column *scope* specifies which aspect – diagnosis, care, sensory – of ASD was investigated by the respective authors. *Logic support* corresponds both to the non-classical logics that are supported and to specific technologies used alongside them; for example, classical + Machine Learning (ML)/Multi-Criteria Decision Making (MCDM) denotes that the model exclusively supports formal ontologies grounded in first-order logic, combined with machine learning ML and multi-criteria decision making MCDM, but does not include any non-classical logical component. The *hybrid integration* shows, when applicable, the integrations with which the model is compatible. *Clinical applicability* specifies the practical use cases that benefit the most from the model. The *implementation maturity* parameter highlights the research's key technical achievements. The *Year* corresponds to the publication or presentation of the results of the work. Finally, the *key contributions* highlight the main advances in knowledge within the respective research domain. Finally, the *key contributions* describe the main contribution of the work to the advancement of knowledge within the respective research domain.

Our comparative analysis reveals that, although previous work has made valuable contributions, none have realized the full, simultaneous integration of the five non-classical logic architectures proposed in this article. Our approach,

Table 1. Related works with ontologies and hybrid systems for ASD/NDD.

Work	Scope	Logic support	Hybrid integration	Clinical applicability	Implementation maturity	Year	Key contribution
[23] (ADAR)	ASD-specific	Classical DL + SWRL	None	Phenotype inference from ADI-R	Protégé; high accuracy	2015	First OWL-based DSM inference
[14] (AutismOnt)	ASD-specific	Classical DL + SWRL	None	Diagnosis and treatment support	Protégé; 676 classes; BioPortal	2022	Comprehensive classical ASD ontology
[4]	ASD diagnosis	Classical + ML/MCDM	Partial (ML)	Hybrid diagnosis with sociodemographic data	ML + ontology	2022	Machine-learning hybridization
[33] (AsdKB)	ASD (Chinese)	Classical + factual KB	None	Screening and recommendation	RDF/Linked Data	2023	Multilingual knowledge base
[18]	ASD care	Classical + RBR + CBR + process mining	Partial (reasoning)	Hybrid reasoning for care planning	Ontology + CBR	2025	Hybrid ASD care model
[2]	ASD sensory	Fuzzy + temporal	Partial (fuzzy-temporal)	Sensory assessment	Fuzzy-neural framework	2025	Fuzzy sensory framework
HML-AutismOnt	ASD/NDD	Fuzzy + paraconsistent (Eτ) + temporal + modal + deontic	Full	Longitudinal decision support; explainable and ethically grounded	Protégé + custom hybrid reasoner	2026	First five-logic hybrid framework

therefore, marks a decisive advance in meeting the complex demands of modeling real-world scenarios of ASD.

3 Proposed Hybrid Multi-logic AutismOnt Framework

This section introduces the main technical contribution of this work: HML-AutismOnt, a hybrid multi-logic framework that extends the established AutismOnt ontology through a modular, layered architecture. This architecture preserves full backward compatibility with existing OWL 2 DL–based applications, while supporting the seamless integration of five complementary non-classical logics—modal, temporal, fuzzy, deontic, and paraconsistent — facilitating meaningful reasoning in ASD.

3.1 Architectural Overview

The framework is implemented using a modular, layered architecture with four layers, as shown in Fig. 1.

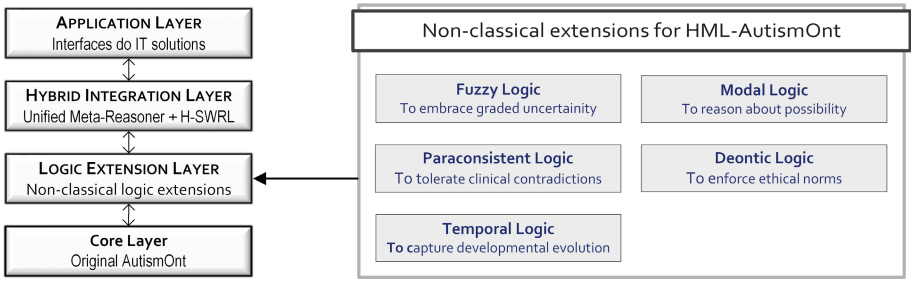


Fig. 1. HML-AutismOnt: Four-Layer Hybrid Multi-Logic Framework Architecture.

This layered design ensures that legacy applications that depend on the original AutismOnt continue to function without modification, while new hybrid capabilities become immediately available. In the following, we describe each layer and the meta-reasoner algorithm responsible for delegating and synthesizing H-SWRL[1] rules or hybrid queries.

The HML-AutismOnt framework comprises four layers. The ***Core Layer*** hosts the original AutismOnt ontology [14], preserving all 676 classes, 124 properties, and the existing SWRLs rules. The ***Logic Extension Layer*** implements five independent non-classical logic extensions—one for each logic—through annotation properties and custom reasoners: fuzzy, paraconsistent, temporal, modal, and deontic. The ***Hybrid Integration Layer*** consists of a unified meta-reasoner and the extended rule language H-SWRL, which allows atoms from different logics to be combined within a single rule. Finally, the ***Application Layer*** provides interfaces for clinical decision support, longitudinal prediction, explainability, and integration with electronic health records.

In fact, as clinical questions or patient assertions enter the ***Application Layer*** and are formalized as RDFs triples with hybrid reasoning goals, the request is forwarded by standardized Application Programming Interface (APIs) to the Hybrid Integration Layer. The Meta-Reasoner parses the incoming H-SWRL rule or query, decomposes it into classical OWL 2 DL and non-classical atoms, constructs an enriched ontology snapshot, and dispatches sub-tasks to the ***Logic Extension Layer***. The five extensions operate concurrently and return partial results as annotated triples, certainty values, temporal intervals, modal

[1] H-SWRL: A syntactic extension of standard SWRL that supports mixed atoms – classical, fuzzy-graded, paraconsistent-evidential, temporal, modal, and deontic – in the same body/head of rule. Parsing and translation to executable form are performed at runtime via a parser.

labels, and normative assessments to the Meta-Reasoner, which integrates these results into a unified hybrid inference with full provenance and returns them to the ***Application Layer*** for clinical output generation.

3.2 Core Layer

The ***Core Layer*** constitutes the immutable foundation of the HML-AutismOnt framework. It hosts the original AutismOnt ontology [14], encoded natively in OWL 2 DL, preserving all 676 classes, 124 object and data properties, and the complete set of pre-existing SWRL rules. The layer supports classical reasoning exclusively through standard OWL 2 DL reasoners such as HermiT, Pellet, and FaCT++. The diagram in Fig. 2 illustrates the internal structure and interfaces of the Core Layer.

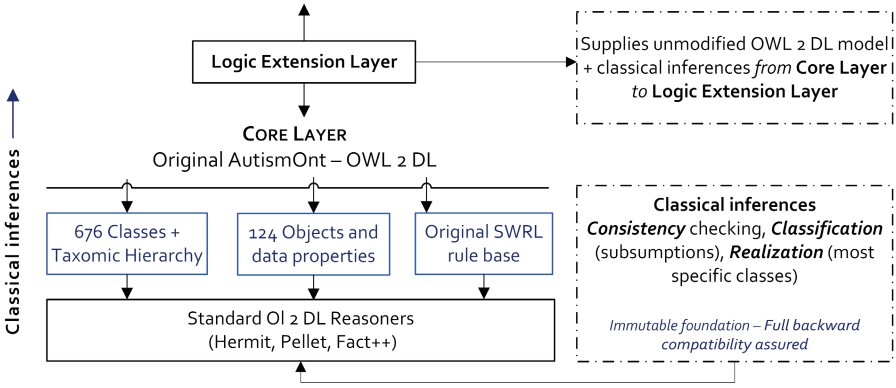


Fig. 2. ***Core Layer*** of the HML-AutismOnt.

The Core Layer supplies the entire unmodified base ontology, together with the results of the classical inferences above, to the logic extension layer via standard loading of the OWL API model.

The main classical inference services provided by the ***Core Layer*** are *consistency checking*, *classification*, and *realization*.

Consistency Checking: Verifies whether the ontology (TBox + RBox + ABox + CBox) is satisfiable and contains no logical contradictions. Real-world example: If an individual patient is asserted to belong simultaneously to the disjoint classes `SevereAutism` and `NoSymptoms`, the reasoner immediately detects an inconsistency and prevents any further inference.

Classification computes the complete inferred class hierarchy by deriving all subsumption relations ($\sqsubseteq$). Real-world example: The reasoner automatically infers that `SocialCommunicationDeficit` $\sqsubseteq$ `CoreASDPhenotype` and that `HighFunctioningAutism` $\sqsubseteq$ `AutismSpectrumDisorder`, even when these relations are not explicitly asserted.

Realization determines, for each individual, the most specific (most restrictive) classes to which it belongs. Real-world example: Given the patient instance `Alex` with observed properties, the reasoner realizes that `Alex` is an instance of `ModerateASD`, rather than merely the more general `AutismSpectrumDisorder`.

In runtime, the ***Logic Extension Layer*** receives the unmodified OWL 2 DL model from the ***Core Layer*** as input and applies non-classical extensions only as annotation overlays and external reasoners. This strict compartmentalization ensures that legacy applications can continue to issue queries directly to the ***Core Layer*** and receive exclusively classical inferences, without any requirement to recognize or interact with the hybrid extensions implemented in the layers above it.

3.3 Logic Extension Layer

The ***Logic Extension Layer*** implements five independent, pluggable extensions—one for each non-classical logic. To optimize implementation straightforwardness and full backward compatibility, every extension is realized exclusively through OWL annotation properties, added at runtime via the OWL API, and lightweight custom reasoners or SWRLs built-ins. We made no alterations to the axioms of the ***Core Layer***; the extensions operate as optional superimposed modules. This design preserves full modularity, allowing each extension to be activated or deactivated independently at deployment time. We incorporated the logic as follows:

- **Fuzzy Logic** through the FuzzyOWL 2 plugin, annotating graded properties such as `hasSocialCommunicationDeficitSeverity` and `hasSensoryHypersensitivityLevel` aligned with ASD severity levels.
- **Paraconsistent** Annotated Evidential Logic Eτ by annotating assertions with pairs of favorable and contrary evidence (μ, λ), supported by a custom lightweight reasoner that computes net certainty without trivialization.
- **Temporal Logic** through Linear Temporal Logic (LTL) operators expressed via annotation properties and custom SWRL built-ins.
- **Modal Logic** using Kripke-style possible-world annotations (`modalPossible` and `modalNecessary`).
- **Deontic Logic** through annotation properties that express obligation (O), permission (P), and prohibition (F).

The *interface point* directed downward toward the ***Core Layer*** receives unmodified AutismOnt classes, properties, and SWRL rules through standard OWL model loading mechanisms. Conversely, the *interface point* directed upward toward the ***Hybrid Integration Layer*** returns an enriched, annotated ontology snapshot, accompanied by partial inference results, such as graded membership values, certainty measures, temporal intervals, modal labels, and normative verdicts.

3.4 Hybrid Integration Layer

The ***Hybrid Integration Layer*** acts as the central orchestration engine of HML-AutismOnt. This layer is stateless and fully modular, allowing any subset of extensions to be activated without changes to the code. It unifies classical OWL 2 DL reasoning with the five non-classical extensions provided by the ***Logic Extension Layer***, enabling atoms from different logics to be combined within a single rule while preserving full backward compatibility. For maximum ease of implementation, the layer can be implemented entirely on top of the OWL API (Java). It consists of the two tightly coupled, self-contained components: *i*) *Unified Meta-Reasoner* that is a lightweight orchestrator that *i*) parses incoming H-SWRL rules/queries, *ii*) decomposes each atom by logic type, *iii*) dispatches sub-tasks in parallel to the appropriate extensions in the ***Logic Extension Layer***, *iv*) collects partial results, and *v*) performs result synthesis, consistency checking, and provenance tracking. The Unified Meta-Reasoner pseudo-code algorithm is provided in Algorithm 1.

The associated interface points originate downward from the ***Logic Extension Layer***, which obtains enriched annotations and intermediate inference outputs (including graded memberships, certainty levels, temporal intervals, modal labels, and normative verdicts). They also originate indirectly from the ***Core Layer***, which is accessed via the Logic Extension Layer to reach the unaltered AutismOnt base model. On the upward side, the interface points connect to the ***Application Layer***, which delivers consolidated hybrid inference results along with complete provenance information and explanation traces.

3.5 Application Layer

The Application Layer constitutes the outermost tier of HML-AutismOnt. The layer is deliberately thin and stateless, performing no reasoning itself and ensuring complete separation of concerns while preserving full usability and interoperability. Consume consolidated hybrid inferences produced by the Hybrid Integration Layer and expose them through user-facing standardized interfaces for clinical decision support, longitudinal prediction, explainability, and integration of Electronic Health Record (EHR). For maximum ease of implementation, the layer can be realized as a set of lightweight standards-compliant components built on common web technologies, such as Spring Boot, React, and FHIR. The main components accommodated in this layer include the following:

- *Clinical Decision Support Module*: REpresentational State Transfer (REST)-full API endpoints that deliver real-time assessment of the ASD profile, severity evaluations, and intervention recommendations.
- *Longitudinal Prediction Module*: Web service that generates developmental trajectories and future-state forecasts using temporal and fuzzy inferences.
- *Explainability & Provenance Module*: Generates human-readable reports and visualizations that include graded certainty, temporal context, modal possibilities, normative justifications, and full inference provenance.

– *Legacy Integration Service*: Bidirectional endpoints to import patient data and export hybrid reasoning results to appropriate information systems.

The *interface point* directed downward toward the ***Hybrid Integration Layer*** receives consolidated hybrid inferences, including graded memberships, certainty values, temporal intervals, modal labels, normative verdicts, and complete provenance traces. In contrast, the *interface point* directed outward toward *external actors and systems* exposes REST API, web dashboards, Fast Healthcare Interoperability Resources (FHIR) resources, and export formats to end users, investigators, and interoperable platforms.

3.6 Hybrid Rule Language and Inference

The central innovation is the hybrid extension H-SWRL, which allows atoms from all five logics to appear in a single rule. Three representative rules illustrate the approach.

Rule 1: Social Communication and Intervention Recommendation

$$\frac{\begin{array}{l}\mathsf{Patient}(p)\\ \mathsf{Fuzzy}(p, \mathit{SocCommDef}, \mathit{Moderate}, \delta \geq 0.65)\\ \mathsf{Para}(p, \mathit{RRB}, \mu \geq 0.60, \lambda \leq 0.35)\\ \mathsf{Temp}(p, \mathbf{F}_{\leq 12\mathrm{m}}(\mathit{ABA} \rightarrow \mathit{SocImp} \geq 0.30))\\ \mathsf{Modal}(p, \Diamond\, \mathit{ComorbidAnxiety})\\ \mathsf{Deon}(\mathit{Recommend}(p, \mathit{HighIntensityABA}))\end{array}}{\mathsf{Infer}(p, \mathit{HighPriorityABA_Monitoring}) \wedge \mathsf{Cert}(0.82)} \quad (1)$$

Rule 1 means that if a patient exhibits a moderate deficit in social communication with a fuzzy confidence level of at least 0.65, presents restricted and repetitive behaviors parameterized by $\mu \geq 0.60$ and $\lambda \leq 0.35$, has demonstrated a social improvement of at least 0.30 following Applied Behavior Analysis (ABA) intervention within the past 12 months, may have comorbid anxiety, and high-intensity ABA is clinically recommended, then it is inferred that the patient requires high-priority ABA monitoring, with a certainty of 0.82.

Rule 2: Sensory Sensitivity and Early Intervention Window

$$\frac{\begin{array}{l}\mathsf{Patient}(p)\\ \mathsf{Fuzzy}(p, \mathit{SensoryHypersensitivity}, \mathit{High}, \delta \geq 0.75)\\ \mathsf{Para}(p, \mathit{SleepDisturbance}, \mu \geq 0.70, \lambda \leq 0.25)\\ \mathsf{Temp}(p, \mathbf{G}_{0\text{–}36\mathrm{m}}(\mathit{EarlySensoryIntegrationTherapy} \rightarrow \mathit{ReducedMeltdowns}))\\ \mathsf{Modal}(p, \Box\, \neg \mathit{SevereLanguageDelay})\\ \mathsf{Deon}(\mathit{Recommend}(p, \mathit{SensoryIntegrationTherapy}))\end{array}}{\mathsf{Infer}(p, \mathit{UrgentSensoryIntegrationProtocol}) \wedge \mathsf{Cert}(0.79)} \quad (2)$$

The rule 2 means that if a patient exhibits high sensory hypersensitivity with a fuzzy confidence level of at least 0.75, presents a sleep disturbance parameterized by $\mu \geq 0.70$ and $\lambda \leq 0.25$, has shown reduced meltdowns after early sensory integration therapy in the window of 0 to 36-months, does not have severe language delay, and sensory integration therapy is clinically recommended, then it is inferred that the patient requires an urgent sensory integration protocol, with a certainty of 0.79.

Rule 3: Comorbidity Risk and Ethical Shared Decision-Making

$$\frac{\begin{array}{l}\mathsf{Patient}(p)\\ \mathsf{Fuzzy}(p, \mathit{IntellectualDisabilityRisk}, \mathit{Moderate}, \delta \geq 0.55)\\ \mathsf{Para}(p, \mathit{EpilepsyComorbidity}, \mu \geq 0.65, \lambda \leq 0.30)\\ \mathsf{Temp}(p, \mathbf{U}_{\leq 8\mathrm{y}}(\mathit{RegularEEGMonitoring} \rightarrow \mathit{EarlyEpilepsyDetection}))\\ \mathsf{Modal}(p, \Diamond\, \mathit{IntellectualDisability})\\ \mathsf{Deon}(\mathit{Permit}(\mathit{SharedDecisionMakingWithParents}(p, \mathit{GeneticTesting})))\end{array}}{\mathsf{Infer}(p, \mathit{ConditionalGeneticTestingProtocol}) \wedge \mathsf{Cert}(0.76)} \quad (3)$$

Finally, rule 3 means that if a patient exhibits a moderate risk of intellectual disability with a fuzzy confidence level of at least 0.55, presents an epilepsy comorbidity parameterized by $\mu \geq 0.65$ and $\lambda \leq 0.30$, has demonstrated early detection of epilepsy resulting from regular EEG monitoring up to 8 years of age, may have intellectual disability, and shared decision-making with parents regarding genetic testing is ethically permitted, then it is inferred that the patient requires a conditional genetic testing protocol, with a certainty of 0.76.

3.7 The Meta-Reasoner Delegation Algorithm

To provide a concrete example of the orchestration mechanism, Algorithm 1 fully specifies the delegation and synthesis process executed by the Unified Meta-Reasoner.

The algorithm guarantees complete traceability by linking each inference step to its origin logic extension and maintains a strict separation of concerns between orchestration and specialized reasoning.

Algorithm 1. Unified Meta-Reasoner Delegation and Synthesis Algorithm

Input: H-SWRL rule R or hybrid query Q
Output: Consolidated hybrid inference result I_{hybrid} with provenance trace P
$A \leftarrow ParseR, Q$
$A = \{a_1, a_2, \ldots, a_n\}$
    $a_i \in A$
$T(a_i) \leftarrow IdentifyLogicTypea_i$
$T(a_i) \in \{\text{Classical}, \text{Fuzzy}, \text{Paraconsistent}, \text{Temporal}, \text{Modal}, \text{Deontic}\}$
$A_{\text{classical}} \leftarrow \{a_i \in A \mid T(a_i) = \text{Classical}\}$
$A_{\text{nonclassical}} \leftarrow A \setminus A_{\text{classical}}$
Delegation phase
$\mathcal{R}_{\text{nc}} \leftarrow \emptyset$
    $a_i \in A_{\text{nonclassical}}$
$E_i \leftarrow E_{T(a_i)}$
$r_i \leftarrow ReasonE_i, a_i$
$\mathcal{R}_{\text{nc}} \leftarrow \mathcal{R}_{\text{nc}} \cup \{r_i\}$
Classical reasoning phase
$\mathcal{R}_{\text{cl}} \leftarrow OWLReasoningA_{\text{classical}}$
$\mathcal{R}_{\text{partial}} \leftarrow \mathcal{R}_{\text{cl}} \cup \mathcal{R}_{\text{nc}}$
Synthesis phase
$\Delta_{\text{fuzzy}} \leftarrow AggregateFuzzyDegrees\mathcal{R}_{\text{partial}}$
$\Theta_{\text{temporal}} \leftarrow PropagateTemporalRelations\mathcal{R}_{\text{partial}}$
$\Phi_{\text{modal}} \leftarrow ApplyModalFilters\mathcal{R}_{\text{partial}}$
$\Omega_{\text{deontic}} \leftarrow ApplyDeonticConstraints\mathcal{R}_{\text{partial}}$
$C_{\text{para}} \leftarrow ComputeParaconsistentCertainty\mathcal{R}_{\text{partial}}$
$C_{\text{para}} = \mu - \lambda$
$\mathcal{F} \leftarrow DetectCrossLogicConflicts\mathcal{R}_{\text{partial}}$
$I_{\text{hybrid}} \leftarrow IntegrateResults\Delta_{\text{fuzzy}}, \Theta_{\text{temporal}}, \Phi_{\text{modal}}, \Omega_{\text{deontic}}, C_{\text{para}}, \mathcal{F}$
$P \leftarrow BuildProvenanceTraceA, \mathcal{R}_{\text{partial}}, I_{\text{hybrid}}$
return (I_{hybrid}, P)

4 Illustrative Example: Longitudinal Decision-Support Scenario

To demonstrate the practical utility of the HML-AutismOnt framework, we present a longitudinal decision-support scenario involving a representative pediatric case. The example spans an 18-month period and integrates heterogeneous data sources that capture uncertainty, temporal evolution, and potentially conflicting observations. Figure 3 illustrates the workflow associated with this scenario.

4.1 Input Scenario

We consider the case of Alex (age 6y2m) with heterogeneous input data derived from multiple sources, including developmental history, behavioral observations, and standardized evaluation results. The input indicates delayed language development, difficulties in social interaction, sensory hypersensitivity, and variable

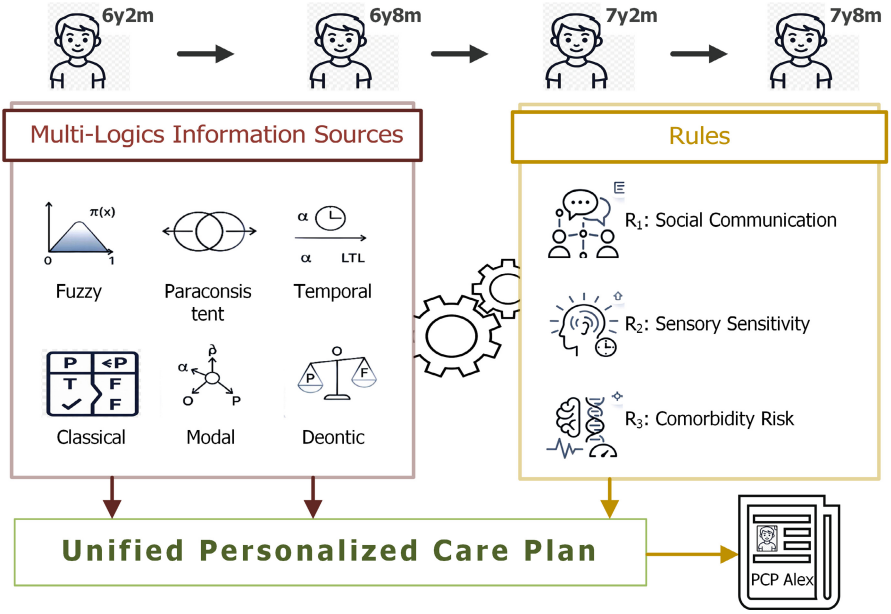


Fig. 3. Workflow of a hypothetical evaluation performed in an environment that implements our architecture.

reports in different contexts. We map this information into structured ontology instances and annotate it across multiple logical dimensions, including graded severity (fuzzy), potentially conflicting observations (paraconsistent), temporal evolution, and contextual constraints.

4.2 Initial Assessment Data (Age 6y2m)

Standard assessment instruments (e.g. Autism Diagnostic Observation Schedule, 2nd Ed. (ADOS-2), ADI-R, acSRS-2) were used to derive structured input observations. Key findings include moderate-to-high social-communication difficulties, evidence of repetitive behaviors, and variability between measures reported by parents and teachers. The key findings of the initial assessment were as follows:

- Moderate impairment in social-communication and repetitive behaviors;
- High severity reported by parents and moderate scores reported by teachers;
- Significant sensory hypersensitivity and sleep disturbances.

Figure 4 illustrates Alex's decision-support scenario milestones through the 18-month time frame.

4.3 Application of Hybrid Rules

The HML-AutismOnt framework processed the initial multi-source data and generated three integrated recommendations.

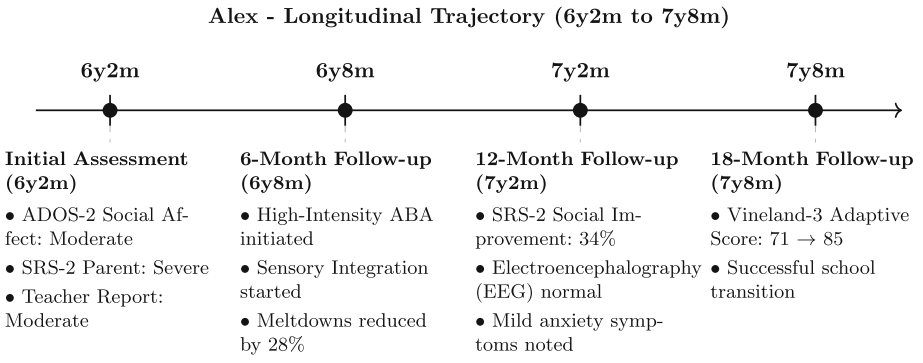


Fig. 4. Longitudinal trajectory of Alex from age 6y2m to 7y8m.

Application of Rule 1—Social Communication and Intervention Recommendation. The framework inferred moderate social-communication difficulties combined with net positive evidence for restricted and repetitive behaviors (despite some differences between home and school reports). Considering the expected developmental window and the possibility of emerging anxiety, the framework recommended a high-intensity applied behavior analysis ABA with regular monitoring (25–30 h per week).

Application of Rule 2—Sensory Sensitivity and Early Intervention Window. High sensory hypersensitivity and associated sleep difficulties were detected. The framework highlighted the importance of acting within the remaining critical developmental period and recommended daily sensory integration therapy (45-minute sessions).

Application of Rule 3—Comorbidity Risk and Ethical Shared Decision-Making. A moderate risk of intellectual disability was identified, along with a possible comorbidity of epilepsy. The framework produced a conditional recommendation for genetic testing (chromosomal microarray and epilepsy gene panel) only after a shared decision-making conversation with Alex's parents, while recommending regular EEG monitoring until age 8.

4.4 Longitudinal Evolution

The clinical team implemented the recommendations and monitored progress through regular follow-up appointments. At 6 months (age 6y8m), Alex had begun high-intensity ABA and daily sensory integration sessions. His social score on Social Responsiveness Scale, Second Edition (SRS-2) improved by 19%, and parent-reported meltdowns decreased by 28%. The results of EEG remained normal. The framework suggested a modest intensification of the social skills component within the existing plan.

At 12 months of follow-up (aged 7y2 m), the SRS-2 social improvement was 34%, and the meltdowns had decreased by 47%. Epilepsy was not identified.

Mild anxiety symptoms appeared, consistent with the possibility noted earlier. The framework updated the care plan accordingly.

At 18-months of follow-up (aged 7y8 m), all main goals had been met or exceeded. Alex's overall adaptive functioning (Vineland-3) increased from a standard score of 71 to 85. He successfully transitioned into the second grade with reduced support needs. The framework continued to provide ongoing guidance for the next stage of development.

This example illustrates how the HML-AutismOnt framework integrates heterogeneous, multi-source data into a coherent, adaptive decision-support process, enabling structured reasoning over evolving, potentially inconsistent information.

5 Discussion

In this work, we demonstrate that the HML-AutismOnt framework represents a significant methodological advance in the semantic modeling of neurodevelopmental disorders, with particular emphasis on ASD. By extending the well-established AutismOnt [14] through the concomitant integration of five complementary non-classical logics—modal, temporal, fuzzy, deontic, and paraconsistent—the framework addresses key limitations of classical Description Logic-based ontologies when applied to the complex, dynamic, and often contradictory reality of ASD.

The primary contribution we achieved lies in the ability to produce more nuanced and clinically relevant inferences. The hybrid rules demonstrated in Sect. 4 combine the graded severity of symptoms (fuzzy), the tolerance of conflicting evidence (paraconsistent), explicit temporal windows (temporal), the differentiation between possibility and necessity (modal), and normative/ethical constraints (deontic) within a single reasoning step. This integration enables us to generate personalized recommendations that better reflect the multifaceted nature of ASD presentations. The longitudinal case study of Alex we presented shows how these recommendations can be applied in practice, supporting more informed decision-making over time and facilitating smoother transitions across developmental stages.

Compared with previous work, we believe that the proposed framework offers clear advantages. Foundational ontologies such as ADAR [23] and AutismOnt itself excel at structured representation and basic phenotype inference, but remain limited by the assumptions of classical logic. More recent hybrid approaches incorporate machine learning [5], case-based reasoning [18], or individual non-classical logics [3], yet none achieves the full concomitant integration of the five logics we present here. This unified approach allows HML-AutismOnt to handle the full spectrum of challenges encountered in real-world ASD care—graded severity, temporal evolution, conflicting data, epistemic uncertainty, and ethical obligations—within a single coherent system.

From a clinical perspective, we consider the framework to have the potential to improve both diagnostic precision and intervention planning. The explainable

nature of hybrid rules facilitates shared decision-making with families and supports longitudinal monitoring with clear, measurable goals. In research settings, it can contribute to a more accurate definition of phenotypes, more powerfully designed studies of developmental trajectories, and improved data harmonization between international cohorts.

Several limitations should be acknowledged. First, the computational cost of full hybrid reasoning increases with the number of active logics and the size of the data set; optimization will be important for large-scale deployment. Second, fuzzy membership functions and evidence thresholds currently rely on expert input and may benefit from data-driven refinement in future versions. Third, although the illustrative case study provides an initial demonstration of feasibility, large-scale prospective validation across heterogeneous system configurations remains necessary to rigorously establish real-world performance and generalizability. Finally, successful implementation will require targeted education and training of practitioners, as well as systematic consideration of the ethical, legal, and regulatory dimensions of AI-enabled decision support platforms. A comprehensive quantitative evaluation and benchmarking against alternative reasoning systems remain important directions for future work.

Looking ahead, several promising research directions emerge. Integration with machine-learning pipelines could enable automated parameter optimization and systematic discovery of novel reasoning rules based on a hybrid ontology. Extension of the framework to additional application domains, such as ADHD and intellectual disability, would substantially enhance its generality and scalability. The development of user-centric interfaces and seamless interoperability with existing information systems through FHIR-compliant standards constitutes a key system engineering challenge for practical deployment.

6 Conclusion

In this paper, we propose HML-AutismOnt, a hybrid multi-logic framework that we developed to extend the established AutismOnt ontology by integrating five non-classical logics: modal, temporal, fuzzy, deontic, and paraconsistent. The proposed approach addresses important limitations of classical DL-based ontologies when applied to the complex, dynamic, and often contradictory nature of ASD and other neurodevelopmental disorders.

By combining graded symptom representation, tolerance of conflicting evidence, explicit temporal modeling, differentiation between possibility and necessity, and formalization of normative and ethical constraints, we enable more nuanced, clinically relevant, and longitudinally consistent reasoning. The illustrative case study of Alex demonstrates how these capabilities can be translated into practical, personalized care recommendations that evolve with child development and respond to new information over time.

Compared to the literature, we believe that HML-AutismOnt offers a more comprehensive solution. Although foundational ontologies such as ADAR [23] and AutismOnt [14] provide robust taxonomic structures, and more recent hybrid

systems have incorporated individual non-classical logic or machine-learning elements, none has achieved the full concomitant integration of the five logics we present here. This unified architecture represents a significant step forward in ontology-based modeling for ASD.

The framework offers several system-level benefits. Enhances decision-support processes, enables richer system workflows, and improves longitudinal data integration across heterogeneous information systems. In research environments, it provides a robust foundation for more precise phenotype modeling, scalable trajectory analysis, and advanced data harmonization across large-scale knowledge bases. From a broader perspective, the approach demonstrates how non-classical logics can effectively bridge the gap between formal knowledge representation and the inherent complexity of real-world application domains.

Naturally, challenges remain. Large-scale validation across diverse contexts remains required, as is computational optimization for large datasets, data-driven refinement of fuzzy membership functions and evidence thresholds, and the development of user-centric interfaces. The ethical, legal, and regulatory considerations surrounding AI-supported decision support systems will also require continued attention.

In conclusion, we consider that HML-AutismOnt demonstrates that the principled integration of non-classical logics can yield a more expressive, flexible, and computationally powerful knowledge representation system. By addressing vagueness, temporality, inconsistency, and normativity within a unified hybrid architecture, the framework advances ontology-driven modeling and opens new avenues for scalable and semantically rich information systems in neurodevelopmental care.

Acknowledgements. We would like to thank healthcare professionals and clinical professionals who participated in the focus group discussions and contributed their expertise to refine the proposed ontology. In particular, we acknowledge Jacqueline de Oliveira Veiga Iglesias, Ana Beatriz dos Santos Carvalho, Patrícia Maria Ligório Quintino de Souza Neri, and Márcia Guimarães Oliveira de Souza for their valuable clinical insights and constructive feedback. Their contributions were essential to ensure the clinical relevance and practical applicability of the rules and recommendations presented in this work.

References

1. Abe, J.M.: Paraconsistent logics and applications. In: Nakamatsu, K., Kountchev, R. (eds.) New Approaches in Intelligent Control. ISRL, vol. 107, pp. 273–300. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32168-4_9
2. Abughazalah, N., Alshahrani, M., Almutairi, A., Alqahtani, F., Alzahrani, S.: An intelligent fuzzy-neural framework for autism sensory assessment using hierarchical linguistic modeling and risk-based temporal decision-making. Sci. Rep. (2025). https://doi.org/10.1038/s41598-025-01234-5
3. Abughazalah, N., et al.: An intelligent fuzzy-neural framework for autism sensory assessment using hierarchical linguistic modeling and risk-based temporal decision-making. PMC (2025). https://pmc.ncbi.nlm.nih.gov/articles/PMC12488974

4. Alqaysi, M.E., et al.: Diagnosis-based hybridization of multimedical tests and sociodemographic characteristics of autism spectrum disorder using artificial intelligence and machine learning techniques. Diagnostics **12**(7), 1585 (2022). https://doi.org/10.3390/diagnostics12071585
5. Alqaysi, M.E., Albahri, A.S., Hamid, R.A.: Diagnosis-based hybridization of multimedical tests and sociodemographic characteristics of autism spectrum disorder using artificial intelligence and machine learning techniques: a systematic review. Int. J. Telemed. Appl. (2022)
6. Alsaedi, S.M., Alsobhi, A., Bitar, H.: Towards an ontology to represent domain knowledge of attention deficit hyperactivity disorder (ADHD): a conceptual model. Int. J. Adv. Comput. Sci. Appl. **15**(12) (2024). https://thesai.org/Publications/ViewPaper?Volume=15&Issue=12&Code=IJACSA&SerialNo=32
7. Artale, A., Kontchakov, R., Kovtunova, A., Ryzhikov, V., Wolter, F., Zakharyaschev, M.: First-order rewritability for ontology-mediated queries in LTL. In: Proceedings of KR 2020: 17th International Conference on Principles of Knowledge Representation and Reasoning, pp. 61–71 (2020)
8. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders, 5th edn., Text Revision (DSM-5-TR). American Psychiatric Publishing (2022)
9. Brandsen, S., Laurenti, L., Logan, R., Richardson, E., et al.: Guiding principles and research priorities for early intervention in autism: Individualized and context-sensitive approaches. Autism Dev. Lang. Impairments **10** (2025). https://doi.org/10.1177/27546330251338419, emphasizes the need for individualized adaptations that vary by context, intensity, and evidence-based practices, aligning directly with DSM-5-TR recommendations for personalized support in ASD
10. Bukiewicz, M.: Paraconsistent foundations of ontological density. PhilArchive (2026). https://philarchive.org/rec/BUKPFO
11. Combi, C., Keravnou-Papailiou, E., Shahar, Y.: Temporal Information Systems in Medicine. Springer, Cham (2010). https://doi.org/10.1007/978-1-4419-6543-1
12. Demrozi, F., et al.: Multimodal AI (MMAI) for next-generation healthcare: data domains, algorithms, challenges, and future perspectives. Comput. Biol. Med. (2025). https://doi.org/10.1016/j.compbiomed.2025.108507
13. Senellart, P.: Provenance in databases: principles and applications. In: Krötzsch, M., Stepanova, D. (eds.) Reasoning Web. Explainable Artificial Intelligence. LNCS, vol. 11810, pp. 104–109. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31423-1_3
14. Hassan, M.M., Mokhtar, H.M.O.: Autismont: an ontology-driven decision support for autism diagnosis and treatment. Egyptian Inform. J. **23**(1), 95–103 (2022). https://doi.org/10.1016/j.eij.2021.07.002
15. Idei, H., et al.: Elucidating multifinal and equifinal pathways to developmental disorders by constructing real-world neurorobotic models. Neural Netw. **169**, 57–74 (2024). https://doi.org/10.1016/j.neunet.2023.10.002
16. Janatimasaoom, B.: Comprehensive Ontology Design for Autism Spectrum Disorder. Master's thesis, Lakehead University, Thunder Bay, Canada (2016). https://knowledgecommons.lakeheadu.ca/handle/2453/3780
17. Kamide, N.: An extended description logic for inconsistency-tolerant ontological reasoning with sequential information. CEUR-WS.org (2022). https://www.scitepress.org/Papers/2020/88764/88764.pdf
18. Kaoura, G., Kovas, K., Boutsinas, B.: A new hybrid reasoning model based on rules, cases and processes: application to care of individuals facing autism spectrum disorders. Knowl. Inf. Syst. **67**, 371–401 (2025). https://doi.org/10.1007/s10115-024-02228-x

19. Letalle, L., et al.: Ontology for mobility of people with intellectual disability: building a basis of definitions for the development of navigation aid systems. In: Krömker, H. (ed.) HCII 2020. LNCS, vol. 12212, pp. 322–334. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50523-3_23
20. Li, X., et al.: Temporal ensemble logic for integrative representation of the entirety of clinical trials. PMC (2025). https://pmc.ncbi.nlm.nih.gov/articles/PMC12875094
21. Lindberg, D.A.B., Humphreys, B.L., McCray, A.T.: The unified medical language system. Methods Inf. Med. **32**(4), 281–291 (1993). https://doi.org/10.1055/s-0038-1634945
22. Michel, C., Ruhrmann, S., Schimmelmann, B.G., Klosterkötter, J., Schultze-Lutter, F.: Course of clinical high-risk states for psychosis beyond conversion. Eur. Arch. Psychiatry Clin. Neurosci. **268**(1), 39–48 (2017). https://doi.org/10.1007/s00406-016-0764-8
23. Mugzach, O., Peleg, M., Bagley, S.C., Guter, S.J., Cook, E.H., Altman, R.B.: An ontology for autism spectrum disorder (ASD) to infer ASD phenotypes from autism diagnostic interview-revised data. J. Biomed. Inform. **56**, 26–36 (2015). https://doi.org/10.1016/j.jbi.2015.06.026
24. Okafor, N., et al.: Hybrid intelligent model for prediction of autism spectrum disorder in children. Preprints.org (2025). https://www.preprints.org/manuscript/202511.0928/v1
25. Organization, W.H.: Package of interventions for rehabilitation: module 5: neurodevelopmental disorders. Technical report, World Health Organization (2023). https://www.who.int/publications/i/item/9789240071193
26. Pham, T., et al.: Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. PMC (2025). https://pmc.ncbi.nlm.nih.gov/articles/PMC12076083
27. Rahayu, N.W., Ferdiana, R., Kusumawardani, S.S.: A systematic review of ontology use in e-learning for students with learning disabilities including dyslexia. Comput. Educ. Artif. Intell. **3**, 100050 (2022). https://doi.org/10.1016/j.caeai.2022.100050
28. Rector, A.L., Rogers, J.E., Pole, P.: The Galen high level ontology. In: Medical Informatics Europe '96, pp. 174–178. IOS Press (1996)
29. Roberts, M.Y., et al.: One size does not fit all for parent-mediated autism interventions: a randomized clinical trial. Autism **27**(2), 443–455 (2023). https://doi.org/10.1177/13623613221102736
30. Tambone, V., et al.: Technological advances in healthcare and medical deontology: towards a hybrid clinical methodology. Healthcare **13**(14), 1665 (2025). https://doi.org/10.3390/healthcare13141665
31. World Health Organization: International classification of diseases 11th revision (ICD-11): autism spectrum disorder. https://icd.who.int/browse11/l-m/en#/6A02 (2019). https://doi.org/10.1002/9781119547167. Accessed 10 Apr 2026
32. Wu, T., et al.: ASDKB: a Chinese knowledge base for the early screening and diagnosis of autism spectrum disorder. Knowl.-Based Syst. **265**, 110–125 (2023)
33. Wu, T., et al.: AsdKB: a Chinese knowledge base for autism spectrum disorder. arXiv preprint arXiv:2307.16773 (2023). https://doi.org/10.48550/arXiv.2307.16773
34. Zhou, S., et al.: Explainable differential diagnosis with dual-inference large language models. PMC (2025). https://pmc.ncbi.nlm.nih.gov/articles/PMC12021655

The Gorizia Railway Terminal as a Strategic Hub: A Study of Logistics Dynamics in Friuli-Venezia Giulia

Giuseppe Borruso(✉), Francesca Sinatra, and Saleh Samimi Dehkordi

University of Tieste, Via Alfonso Valerio 4/1, Trieste, Italy
giuseppe.borruso@deams.units.it, {francesca.sinatra, saleh.samimidehkordi}@phd.units.it

Abstract. The initiative stems from a desire to assess the feasibility of reviving logistics activities in the area of Gorizia, which is considered a key asset for the development of the region and the local community. The Gorizia Railway terminal could, in fact, play a crucial role in supporting supply chains and integrating different modes of transport, helping to reduce logistics costs and environmental impact. The project outcomes fall into two distinct categories. The first stage involved obtaining as detailed an overview as possible of the current situation. The project aims to provide a detailed analysis of freight transport flows, focusing on the different categories of goods and the modes of transport used. Furthermore, the project aims to identify the stakeholders involved in transit traffic through the Gorizia freight village, taking into account the main road and rail operators, and narrowing the scope of the analysis to operational, managerial and infrastructural aspects (e.g. sites, warehouses, railway tracks, manoeuvring yards, goods handling equipment, etc.), highlighting the strengths and weaknesses of the regional infrastructure.

Keywords: Gorizia Dry Port · Railway Terminal · Logistics

1 Introduction

The project formed part of a scenario analysis concerning the development of the Gorizia Dry port's intermodal terminal, in light of the key challenges posed by the achievement of the SDGs relating to the social, economic and environmental sustainability of the transport sector, including through innovative integrated approaches, in line with the pillars of Spoke 8 of the iNEST project (focused on maritime technologies, sustainability and innovation) [1].

iNEST—Interconnected Nord-Est Innovation Ecosystem—is one of the 11 Innovation Ecosystems funded by the PNRR–NextGenerationEU. The activities of Spoke 8 are organised into five strategic areas, including sustainable maritime and inland waterway transport and integrated sea-land maritime planning, which are naturally extended to the challenges of multimodal logistics and the interconnection between port and hinterland hubs such as the one in Gorizia.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 446–455, 2027.
https://doi.org/10.1007/978-3-032-30494-0_28

The European regulatory framework further reinforces the project's relevance. The European Commission's Strategy for Sustainable and Smart Mobility sets the target of increasing rail freight volumes by 50% by 2030 and doubling them by 2050. The shift from long-distance road transport to rail for distances exceeding 700 km could save 40 million tonnes of CO_2 per year, bringing the modal share of rail freight transport to 36% in the EU. In this context, the development of intermodal hubs such as the Gorizia Interport becomes a direct means of implementing the sustainability objectives of the 2030 Agenda, in particular SDG 9 (resilient infrastructure and sustainable industrialisation), SDG 11 (sustainable cities and communities) and SDG 13 (climate action) [1].

The freight village is a crucial hub for freight transport and international logistics, given its strategic proximity to the Slovenian border, within the broader context of increasing freight traffic to and from Central and Eastern Europe, combined with the need for a modal shift from road to rail and the relocation of logistics and transport functions to dry ports, with a corresponding expansion of the terminal's infrastructure and operational components [2].

The benefits of this analysis related to the following aspects:

- economic: growth in the market segment, new jobs and competitive advantages;
- social, in terms of regional development and improved quality of life;
- environmental: an efficient logistics system contributes to the reduction of pollutant emissions.

Another aspect concerned the potential development pathways for the Isonzo area, situated within the network of port and hinterland infrastructure hubs in the relevant geographical area. This development should be understood both in terms of its evolution as a logistics hub within the wider freight transport system and with regard to the potential for infrastructural expansion, taking into account both existing constraints arising from planning and programming instruments, as well as expansion constraints linked to the presence of the international border and the urban context of the Isonzo area.

Alongside these levels of understanding lies the fundamental objective of pursuing the transition towards sustainable transport and logistics systems, aimed at minimising environmental impacts, which can also be achieved by encouraging a modal shift from road to rail transport. In this regard, therefore, particular attention was paid to the developmental aspects linked to the infrastructure enhancement of the hub, both in terms of specific infrastructure and its linear components.

2 Context

The Gorizia Interport is an intermodal logistics hub covering an area of 600,000 square metres, offering a wide range of services for international trade and road and rail transport to and from countries in Central and Eastern Europe. The area under consideration covers a total surface area of approximately 4,000 square metres and includes a railway terminal, currently in a state of disuse, located within the Gorizia freight village. It contains a 3,000-square-metre industrial warehouse, as well as a large area connected by railway tracks, which gives the site significant strategic potential from a logistical and infrastructural perspective.

This study stems from a desire to assess the feasibility of reactivating logistical activities in the area, which is considered a key asset for the development of the region and the local community. The rail terminal could, in fact, play a crucial role in supporting production chains and integrating transport modes, helping to reduce logistics costs and environmental impact.

The aim of the study is twofold: on the one hand, to understand the type and volumes of goods currently transiting through the Gorizia area and, on the other, to map the production activities present in the area and their associated logistics requirements.

To achieve these objectives, it is essential to identify the main transport operators working in the Gorizia freight village area, analysing their flows, the origin and destination of goods, as well as the predominant modes of transport (road, rail, intermodal).

3 Metodology

The innovative aspect of the project's implementation involves structuring the analysis and design proposal through the implementation of a Spatial Decision Support System (SDSS), integrated into the development of 'As is' and 'To be' scenarios.

An SDSS is an IT tool designed to support decision-making in contexts with a significant geographical component, integrating geographical data, models and analytical tools to support the evaluation and selection of options. This is combined with the conceptualisation of 'As is' (existing scenario) – 'To be' (desired scenario) scenarios, mapping out the processes and operations of the logistics hub in line with future developments in transport flows at the Gorizia dry port [3] (Figs. 1 and 2).

Multiscale analysis is an approach that allows the system to be analysed across different levels and scenarios. One level of analysis involves the classification and observation of phenomena at the 'macro' and 'regional' scales (Friuli-Venezia Giulia, extending to the neighbouring areas of Slovenia, Veneto and Austria): e.g. the current state of linear and point infrastructure; flows within the reference territory; the territory's capacity to generate/attract flows. A second level involves consideration of the 'micro' level, namely the Gorizia dry port and its territorial area. This concerns the characteristics and infrastructure facilities of the hub, as well as the planning instruments within the related context and of interest to the reference context.

The levels of observation refer to two points in time: 'As is' (initial) and 'To Be' (project).

The 'overlay' analysis allows for a multi-criteria weighted overlay to classify the area based on a series of attributes and different layers of geographic data, with the aim of producing various scenario maps that combine the different pieces of information.

Central to this analysis is the acquisition, collection and organisation of the various types of geographically referenced data relevant to the project under study.

The analysis focuses on a geographical buffer zone of approximately 50 km, with the aim of identifying economic operators who may be interested in utilising the services that could be offered by the logistics area, including in collaboration with ARS [4–6].

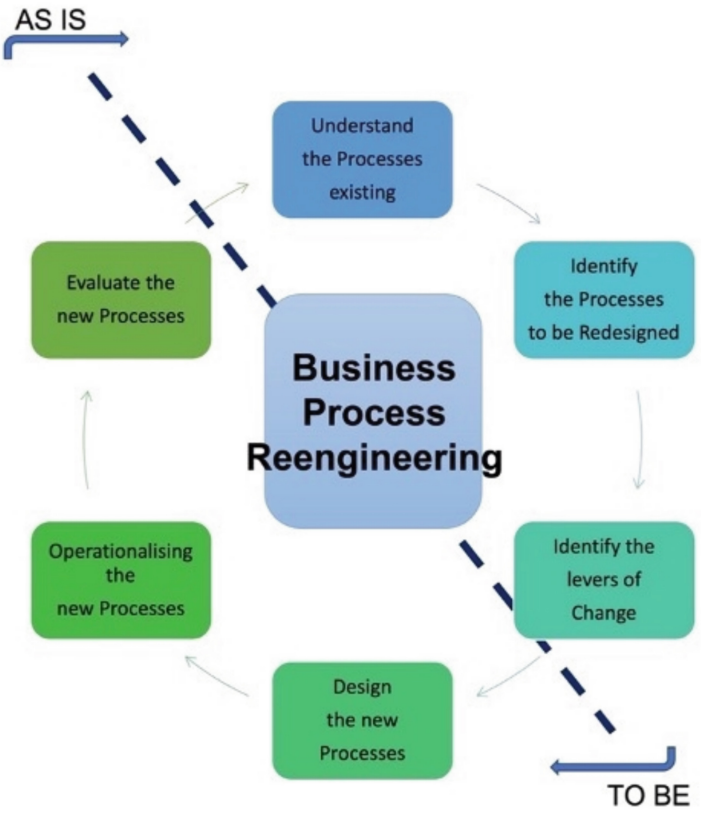


Fig. 1. The framework of standard analytical tools.

The research is divided into several phases, combining qualitative and quantitative approaches:

- Desk research: collection of secondary data and preliminary analysis of the target area, including an economic and infrastructural overview.
- Targeted questionnaires: to analyse operators' logistics and customs requirements; to gather information on routes, modes of transport and perceived critical issues.
- Flow analysis: study of goods movements with particular attention to origin-destination routes, both by road and rail.
- Multi-level survey: distribution of questionnaires to local and regional economic operators to understand the strategic orientation of their activities and their propensity towards intermodality.
- Mapping of local stakeholders: identification and classification of operators potentially interested in using the logistics hub.
- Organisation of thematic days: events and workshops dedicated to promoting the freight village and discussing opportunities related to intermodality, with the involvement of stakeholders [3, 7, 8].

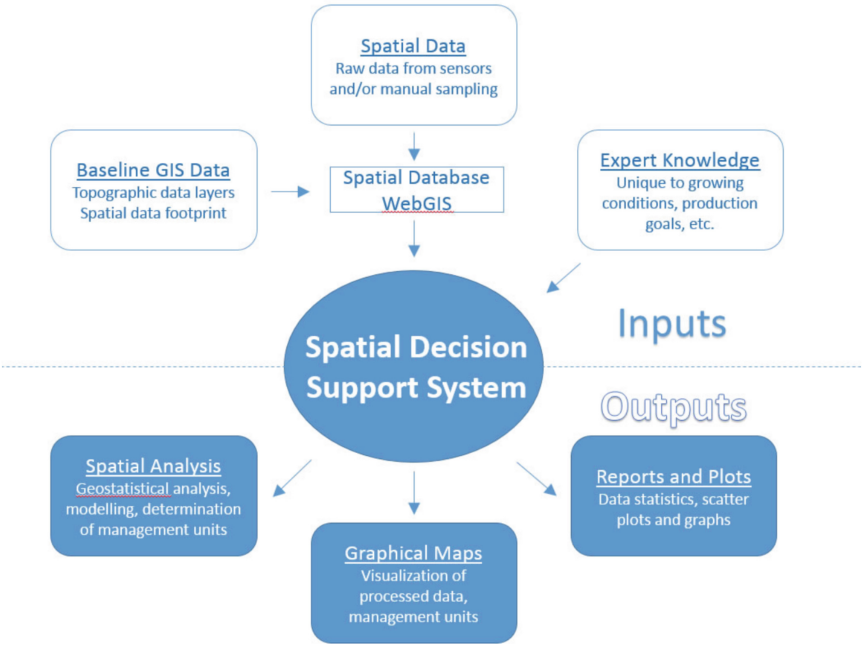


Fig. 2. Sdss: 'As is' - 'To be'

4 Data

4.1 A Gorizia Dry Port Overview

The Gorizia Intermodal Terminal, managed by SDAG (Gorizia dry port), is a major logistics hub located on the border between Italy and Slovenia. It covers an area of over 600,000 square metres and offers a wide range of services, including customs operations, cold storage facilities and infrastructure for intermodal transport. The key components of the infrastructure are as follows:

- Total area: 600,000 m^2
- Intermodal warehouse: a 3,000 m^2 facility with integrated road and rail access.
- Rail infrastructure: five railway tracks equipped to handle intermodal operations.
- Refrigerated storage facilities: 11 cold stores totalling 4,400 m^2, capable of maintaining temperatures down to –30 °C.
- Security measures: the terminal is equipped with CCTV, access control systems and is fully paved, fenced and lit.

Key operational features include:

- Specialisation: it focuses on the logistics of fresh and frozen food products, utilising advanced cold storage capabilities.
- Intermodal services: it offers seamless integration between road and rail transport, improving the efficiency of goods movement.

- Customs services: provides comprehensive customs clearance services, facilitating international trade operations.
- Strategic positioning in cross-border logistics
- Geographical advantage: Located at the crossroads of Italy, Slovenia and Austria, the Gorizia Intermodal Terminal is a key hub in the Trans-European Transport Network (TEN-T). Its proximity to the port of Koper in Slovenia and the port of Trieste in Italy enhances its strategic importance in facilitating trade between Southern and Central Europe.

With regard to cross-border connectivity:

- Rail links: connected to the main rail corridors, it enables efficient freight transport to and from Austria and other Central European countries.
- Road access: direct access to the A34 motorway and other regional roads ensures smooth cross-border road transport operations.
- Collaborative initiatives: involved in various EU-funded projects aimed at improving cross-border transport infrastructure and services, promoting regional economic integration.
- Data on freight flows and destinations: although specific data on freight volumes at the Gorizia intermodal terminal are limited in publicly available sources, regional studies and reports provide detailed information on freight movements and strategic importance:
- Freight consolidation: the terminal helps to consolidate freight flows, improving the profitability of intermodal transport systems through rail services from the terminal to external destinations.
- Integration into supply chains: the terminal's role within the wider logistics network is highlighted in studies focusing on the integration of intermodal terminals into supply chains, emphasising their strategic position in cross-border logistics.

4.2 Geographical and Economic Overview

The area around the SDAG freight village in Gorizia is characterised by a high concentration of logistics, transport and light industrial activities. The concentration of economic activities follows two main axes:

- the west-east axis: along Via dell'Industria and Via Enrico Fermi, which are home to freight forwarding services, logistics companies and manufacturing firms;
- the cross-border axis: leading towards the Slovenian border and Šempeter, with logistics and customs firms serving international traffic.

A well-established logistics-industrial cluster can be observed, thanks to the presence of over 20 companies within a 1 km radius of the freight village, facilitated by its strategic location and proximity to customs crossings. Each of the two areas on the map contains companies classified by sector (e.g. manufacturing, logistics, customs services) (Figs. 3 and 4).

The attached figure provides an overview of the names, sectors and addresses of the companies, including the impact of the consolidation and cross-border integration of the logistics and industrial system.

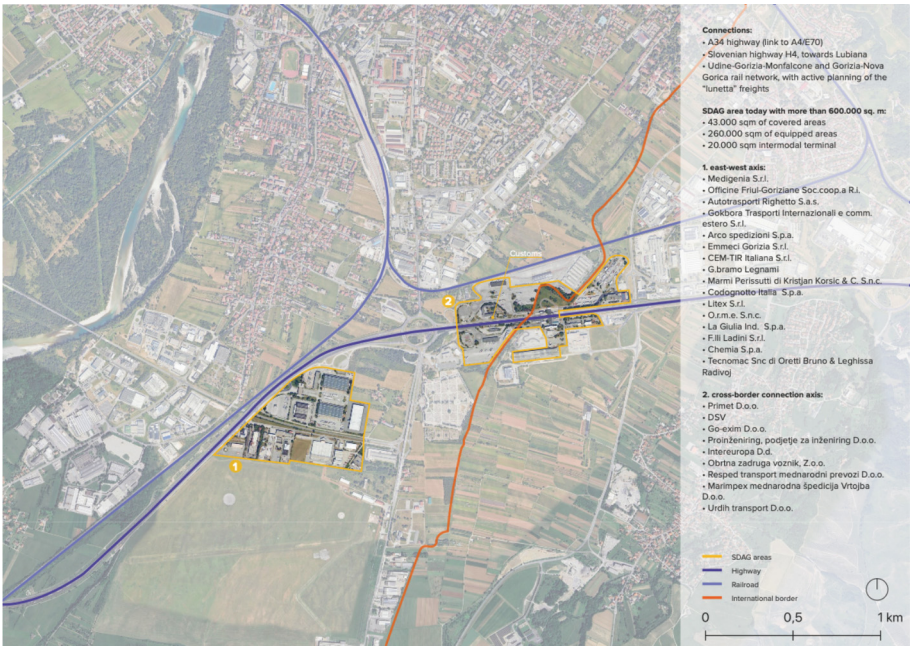


Fig. 3. Geographical and economic overview of the Gorizia dry port.

The €12 million 'rail junction' is currently under construction; it will directly link Udine-Gorizia-Monfalcone to the line for Nova Gorica, eliminating the need to change direction at Gorizia. This will improve connections with the ports of Trieste and Monfalcone.

SDAG has strengthened its partnerships with RFI, the Port System Authority of the Eastern Adriatic Sea and Adriafer to develop intermodal synergies and adopt more sustainable rail operations.

On 25 May 2022, SDAG officially changed its status and obtained the designation of freight village, opening the door to national and European funding.

It has joined the national association UIR (Unione Interporti Riuniti).

Together with the freight villages of Trieste and Cervignano, it presents itself as a single logistics platform with a view to regional synergies: collectively, they offer over 1.5 million m^2 and a rail network with approximately 22 tracks [4–7].

4.3 Database and Questionnaire Construction

To provide an overview of the businesses operating in the Isonzo area, a database has been created, structured and organised using both quantitative and qualitative variables. Its construction was based primarily on unique records such as the company name, supplemented by a geographical variable corresponding to the location of the business, in order to subsequently obtain the georeferencing of the surveyed enterprises and allow for aggregation by territorial area. In addition, variables useful for obtaining a categorical classification were considered, including sectoral affiliation and company size -

Business	Sector	Adress
Medigenia S.r.l.	Pharmaceuticals and pharmacovigilance	Interporto, Gorizia
Officine Friul-Goriziane Soc.coop.a R.i.	Car repair shops and service centers	Interporto, Gorizia
Autotrasporti Righetto S.a.s.	Trucking	Interporto, Gorizia
Gokbora Trasporti Internazionali e comm. estero S.r.l.	International transports	Interporto, Gorizia
Arco spedizioni S.p.a.	Domestic, international transport and integrated logistics	Interporto, Gorizia
Emmeci Gorizia S.r.l.	International shipping and transportation	Interporto, Gorizia
CEM-TIR Italiana S.r.l.	Freight forwarding service	Interporto, Gorizia
G.bramo Legnami	Working lumber	Via dell'Industria 23/25a, Gorizia
Marmi Perissutti di Kristjan Korsic & C. S.n.c.	Funeral monuments	Via dell'Industria 7, Gorizia
Codognotto Italia S.p.a.	Customs forwarders	Via dell'Industria 25, Gorizia
Litex S.r.l.	Rubber hose production	Via dell'Industria 5, Gorizia
O.r.m.e. S.n.c.	Electrical installations, civil and industrial	Via Enrico Fermi 39, Gorizia
La Giulia Ind. S.p.a.	Confectionery production	Via Enrico Fermi 17, Gorizia
F.lli Ladini S.r.l.	Mechanical constructions	Via Enrico Fermi 13, Gorizia
Chemia S.p.a.	Fertilizers	Via Enrico Fermi 33, Gorizia
Tecnomac Snc di Oretti Bruno & Leghissa Radivoj	Wood, machine processing and trade	Via Enrico Fermi 7, Gorizia
Primet D.o.o.	Distribution of meals and raw materials	Mednarodni prehod 2A, Šempeter Pri Gorici
DSV	International logistics service	Mednarodni prehod 2B, Šempeter Pri Gorici
Go-exim D.o.o.	Shipping services	Mednarodni prehod 2B, Šempeter Pri Gorici
Proinženiring, podjetje za inženiring D.o.o.	Engineering consultant	Mednarodni prehod 2B, Šempeter Pri Gorici
Intereuropa D.d.	International logistics service	Mednarodni prehod 8, Šempeter Pri Gorici
Obrtna zadruga voznik, Z.o.o.	Transportation services	Mednarodni prehod 2, Šempeter Pri Gorici
Resped transport mednarodni prevozi D.o.o.	Transportation services	Mednarodni prehod 2, Šempeter Pri Gorici
Marimpex mednarodna špedicija Vrtojba D.o.o.	Customs advisor	Mednarodni prehod 2, Šempeter Pri Gorici
Urdih transport D.o.o.	Transportation services	Mednarodni prehod 2A, Šempeter Pri Gorici

Fig. 4. Significant activities in the proximity of the SDAG area and within their sector

derived from the number of employees - to stratify the sample according to standardised dimensional criteria [8].

To obtain an overview of international trade, two variables were considered: Imports Size and Exports Size. These variables were considered in order to identify potential businesses to which an ad hoc questionnaire could be sent.

The questionnaire was distributed anonymously during the project period to a sample of 142 logistics operators in the sector via the Google Forms platform, which facilitates the processing, analysis and aggregation of results. The aim is to understand how companies currently manage logistics and how the expansion of the terminal could affect economic performance and sustainability. The questionnaire has been divided into several thematic sections. The first part of the questionnaire was designed to gather general information about the companies, whilst the second part was designed to gather information regarding the expected economic and environmental impact associated with the potential development of the terminal. Completing the survey takes approximately 15–20 min. All responses will be treated with the utmost confidentiality and used exclusively for research purposes. As of today, the data collection phase is still ongoing.

5 Conclusion and Future Research

This study is set within a context of strategic importance, focusing on intermodal logistics infrastructure in the context of regional economic development and the transition towards more sustainable freight transport models.

Specifically, the rail terminal located within the Gorizia freight village represents a potential hub for integrating the Friuli Venezia Giulia region into European transport networks. This potential is enhanced by its strategic border location and proximity to the TEN-T corridors. The methodological approach used for the analysis allows for the triangulation of both qualitative and quantitative data, through a combination of territorial analysis, the creation of a company database and the administration of a bespoke questionnaire. The aim of this methodological integration was to obtain a systematic overview of the interdependencies between infrastructure provision and the behavior and perceptions of logistics operators in the region.

Nevertheless, the research is not without limitations, which must be considered to ensure a correct interpretation of the results. Firstly, as data collection is still ongoing, this constitutes a critical variable; meanwhile, the estimation of company value and the scale of trade flows – where not derived from verified official sources – may be subject to inaccuracies due to the self-reported or indirect nature of such information.

In the near future, it would be desirable to expand the sample to include cross-border operators, in order to involve logistics operators from neighboring countries.

Acknowledgments. This study was funded by the "iNEST – Interconnected Nord-Est Innovation Ecosystem" project (ECS_00000043), part of the innovation ecosystem research programme funded by the PNRR – M4C2, Investment 1.5 "Creation and strengthening of innovation ecosystems for sustainability", funded by the European Union – NextGenerationEU (CUP J43C22000320006).

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

References

1. iNEST – Interconnected Nord-Est Innovation Ecosystem Homepage. https://www.consorzioinest.it. Accessed 08 May 2026
2. SDAG S.p.A.: Interporto di Gorizia – Terminal intermodale e servizi logistici. https://www.sdag.it. Accessed 08 May 2026
3. Adriafer Rail Services S.r.l.: GO!Rail – Mobilità ferroviaria sostenibile e integrazione logistica transfrontaliera. https://www.go2025.eu/it/imports/events/proposed/anonymous/go-rail. Accessed 10 Jan 2026
4. Adriafer S.r.l. Homepage. Accessed 20 Feb 2026
5. Adria Ports: Adriafer apre centro manutenzione carri ferroviari a Gorizia. https://www.adriaports.com/it/logistica/adriafer-attiva-centro-manutenzione-carri-ferroviari-a-gorizia/. Accessed 18 Apr 2026
6. European Commission: NextGenerationEU – Recovery Plan for Europe. https://commission.europa.eu/strategy-and-policy/recovery-plan-europe_en. Accessed 07 Apr 2026
7. Regione Autonoma Friuli Venezia Giulia: Piano strategico 2023/2028. https://www.regione.fvg.it/rafvg/cms/RAFVG/GEN/piano-strategico-2023/. Accessed 28 Apr 2026
8. FatturatoItalia. https://www.fatturatoitalia.it/provincia/gorizia. Accessed 07 May 2026

Measuring Demographic Vulnerability and Territorial: a Composite Index Approach

João Lourenço Marques(✉), Jan Wolf, and Joana Duarte

GOVCOPP, Department of Social, Political and Territorial Sciences, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal
jjmarques@ua.pt

Abstract. Low-density territories (LDTs) across Southern Europe have been experiencing persistent demographic contraction, population ageing, and deepening socioeconomic marginalisation. Despite the growing urgency of these challenges for territorial policy, most analytical approaches treat demographic and socioeconomic dimensions in isolation, which limits their capacity to capture the full complexity of territorial decline. This paper proposes an integrated analytical framework that simultaneously assesses demographic vulnerability and territorial attractiveness using a composite index. The framework brings together the Demographic Vulnerability Index (DVI) and a multidimensional Attractiveness Index (AI), the latter structured around three policy-relevant domains: housing conditions, economic performance, and access to general interest services and infrastructure. Methodologically, the approach applies min–max normalisation, polarity adjustment, and a two-stage aggregation procedure. Within each dimension, variables are combined using weighted arithmetic means; across dimensions, a weighted geometric mean is applied, allowing for partial substitutability within domains while penalising severe imbalances across them. When applied to the 165 Portuguese municipalities officially classified as low-density, the framework reveals a pronounced spatial duality. Demographic vulnerability is heavily concentrated in the North–Centre–Alentejo interior corridor, yet territorial attractiveness varies considerably within this same space. The coexistence of structural demographic fragility with localised development assets constitutes one of the central findings of this study. The proposed framework offers a transparent, replicable, and policy-relevant diagnostic tool for differentiating territorial profiles and guiding evidence-based regional interventions.

Keywords: demographic vulnerability · territorial attractiveness · low-density territories · composite indices · Portugal · depopulation · regional policy

1 Introduction

Population decline and demographic ageing have become defining features of low-density territories (LDTs) in Portugal and, more broadly, across Southern and Eastern Europe. What distinguishes these territories is not simply their low population density, but a compounding set of structural processes: demographic contraction, functional fragility, and progressive socioeconomic marginalisation. These dynamics stem

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 456–469, 2027.
https://doi.org/10.1007/978-3-032-30494-0_29

from persistently sub-replacement fertility, selective out-migration that disproportionately removes younger and more educated residents, and spatially uneven economic development. Together, they pose serious challenges to territorial cohesion and have prompted a range of public interventions—from fiscal incentives and employment support to institutional frameworks such as Portugal's Inland Mission Unit (*Unidade de Missão para a Valorização do Interior*).

Assessing the effectiveness of such interventions is methodologically difficult. Policies tend to be diffuse, counterfactual scenarios are hard to construct, and territorial trajectories are highly heterogeneous across municipalities. A further limitation lies in the analytical tools typically used: existing approaches tend to examine either demographic trends or socioeconomic indicators in isolation, thereby preventing a comprehensive understanding of territorial decline. What is needed are integrated frameworks capable of simultaneously capturing structural vulnerability and territorial attractiveness—yet operational models combining these two dimensions in a spatially consistent and policy-interpretable manner remain scarce in the literature.

This paper responds to that gap. It develops a composite index framework that jointly operationalises demographic vulnerability and territorial attractiveness within a single analytical model. The framework was produced under the REVITA project (Strategic Responses for the Revitalisation of Low-Density Territories), within the Science4Policy 2025 programme, and is designed to be methodologically replicable, spatially comparable across municipalities, and directly interpretable for policy purposes. Its empirical application to Portugal's 165 officially classified LDT municipalities demonstrates its diagnostic potential and its capacity to underpin differentiated, evidence-based territorial policy.

2 Theoretical Background

2.1 Depopulation as a Structural Process

Depopulation has gained growing public and academic visibility, often portrayed as an unprecedented contemporary phenomenon. However, as [1] notes, this perception constitutes a 'false novelty': the process is structural and long-standing, not exclusive to the present moment. [2] Coleman and Rowthorn (2011) demonstrate that the fear of population decline is recurrent throughout history, reappearing cyclically and typically amplified by emotional, symbolic, and political readings of demographic data that diverge sharply from analytical interpretations.

[3] characterises contemporary depopulation as a phased structural dynamic. In the initial stage, fertility falls below replacement level, but the natural balance remains positive due to demographic inertia. In a second phase, deepening low fertility and cohort ageing generate a negative natural balance, partially offset by positive migratory flows. Finally, at an advanced stage, even immigration can no longer neutralise population loss, leading to continuous structural decline. Portugal's LDTs have entered this third stage considerably earlier than the national average, making them important laboratories for understanding advanced depopulation dynamics.

2.2 Long-Run Demographic Trends in Portugal

Portugal displays two major national-scale demographic ruptures: in 1970 and 2021. The 1970 rupture resulted from massive post-war emigration to Europe; the 2021 inflexion reflected the simultaneous contraction of natural growth and the migration balance, compounded by the prolonged effects of the 2008 financial crisis. LDTs exhibit a markedly different chronology: structural breaks occurred predominantly between 1950 and 1960, well ahead of national turning points, and decline has been continuous and cumulative ever since [4].

LDTs reached their peak population in 1960 and have since lost more than one-third of their residents, declining from over 3.1 million to below 2 million by 2021 (INE Census data). In contrast, metropolitan municipalities nearly doubled their population over the same period. Expressed as an index (1940 = 100), LDTs stand below 65 while metropolitan areas exceed 200. This structural dualisation between a densely populated coastal strip and a sparsely inhabited interior constitutes the fundamental spatial challenge that the REVITA analytical framework addresses.

2.3 Policy Responses to Territorial Decline

Policy responses to demographic decline have oscillated between two broad strategies: compensatory growth policies centred on attracting population and investment, and adaptation and managed-shrinkage policies that acknowledge irreversibility and seek innovative solutions to maintain quality of life [5, 6]. Evidence consistently shows that most responses remain trapped in a growth culture, denying or ignoring decline and thereby delaying the formulation of more adjusted strategies [7, 8]. Planning for presumed growth aggravates the negative effects of population loss, leading to accumulated infrastructural overcapacity and underutilised equipment.

Emerging concepts such as smart shrinkage [9] and rightsizing [10] advocate for planned management of ‘living with less’, prioritising quality of life, service adequacy, and environmental sustainability over numerical growth targets. For LDTs, differentiated territorial strategies must simultaneously: (a) minimise threats to quality of life during demographic transitions; (b) identify and exploit endogenous opportunities for population attraction; and (c) selectively invest in territorial anchor points with demonstrated capacity for service retention and development diffusion [8].

These strategic imperatives call for analytical tools capable of diagnosing both the depth of demographic vulnerability and the latent potential for territorial attractiveness. This dual requirement informs the composite index framework developed in the following sections.

3 Methodology

3.1 Study Area and Data Sources

This study focuses on the 165 municipalities that Portugal’s 2030 programme formally designates as low-density territories (Deliberation No. 31/2023/PL). In spatial terms, the sample is far from randomly distributed—the vast majority of these municipalities

occupy a contiguous interior corridor spanning the North, Centre, and Alentejo regions, reflecting the well-documented asymmetry between coastal and inland Portugal. Data were drawn from a range of official institutions: INE (censuses, housing statistics, purchasing power, and business accounts), IEFP (unemployment), DGEEC (school equipment), DGS (health services), OpenStreetMap and DGT (road network for accessibility calculations), and the Portuguese Tourism Open Data portal. For each variable, we used the most recent year available. The most recent available data were selected for each variable, producing some temporal heterogeneity across indicators. This small degree of temporal inconsistency across indicators is an unavoidable consequence of this approach, though it was assumed preferable, given that outdated data would undermine the diagnostic value of the framework.

3.2 Composite Index Framework

The analytical framework comprises two complementary composite index families, each capturing a structurally distinct dimension of territorial condition. The first is the Demographic Vulnerability Index (DVI) that characterises each territory's structural exposure to demographic risk and fragility. The second is the Attractiveness Index (AI) that captures multidimensional development conditions through three sub-indices, each corresponding to a central policy domain with direct bearing on territorial fixation capacity: Housing (IHab), Economy (IEcon), and General Interest Services and Accessibility (ISIG). Variables were selected on the basis of: (i) theoretical grounding in the literature on territorial dynamics and demographic decline; (ii) spatial consistency across municipalities, ensuring meaningful comparisons; and iii) policy interpretability, in which, each variable had to be interpretable into concrete intervention priorities, so that weaknesses identified by the index could directly inform policy decisions.

3.3 Normalisation, Polarity Adjustment, and Aggregation

The methodological procedure followed three sequential steps. First, all variables were normalised to a 0–1 scale using min–max normalisation:

$$x_{ij}^{*} = \left(x_{ij} - min_{-j}\right) / \left(max_{-j} - min_{-j}\right) \tag{1}$$

where x_{ij} is the observed value of variable j in municipality i, polarity was then adjusted so that 1 represents the best territorial condition and 0 the worst. Within each dimension, variables were aggregated using a weighted arithmetic mean, reflecting an assumption of partial substitutability among indicators capturing related aspects of the same domain.

By contrast, the overall Attractiveness Index (IA) was computed using a weighted geometric mean across dimensions:

$$I_i = \prod \left(x_{ij}^{*}\right)^{w_j} with \sum w_j = 1 \tag{2}$$

The use of the geometric mean at the aggregate level introduces a non-compensatory structure between dimensions. This choice was preferred over arithmetic alternatives for four reasons: (i) it penalises imbalances across core dimensions, requiring consistent

performance to achieve high index values; (ii) it reflects the multiplicative nature of territorial attractiveness, where housing, economic conditions and services interact rather than combine linearly; (iii) it reduces sensitivity to extreme values; and (iv) it supports a more demanding policy interpretation, whereby critical weaknesses in any dimension constrain overall performance – a desirable property in diagnostics aimed at identifying intervention priorities.

3.4 Demographic Vulnerability Index (DVI)

The DVI integrates four variables capturing complementary structural dimensions of population decline. Table 1 summarises the variables, sources, polarities, and weights. The formula is:

$$IVD_i = 100 \times \left(0.30\,P_{var,i} + 0.20\,P_{ativa,i} + 0.20\,P_{fec,i} + 0.30\,P_{migr,i}\right) \quad (3)$$

Table 1. Variables included in the Demographic Vulnerability Index (DVI).

Variable	Source / Year	Rationale
Resident pop. Change 2011/2021 (%) *(P_VAR$_i$)*	INE, Census 2021	Net result of structural and conjunctural dynamics; persistent decline increases vulnerability
Working-age pop. 15–64 (%) *(P_WORK$_i$)*	INE, Census 2021	Proxy for ageing structure; cumulative and difficult to reverse
Total fertility rate *(P_ISF$_i$)*	INE, 2021	Generational renewal capacity; low values indicate difficulty in self-replacement
Net migration growth rate (%) *(P_MIGR$_i$)*	INE, 2014–2024	Short-term population retention and attraction capacity: a key compensatory mechanism

3.5 Attractiveness Index (AI) and Sub-Indices

The AI aggregates three sub-indices with equal weights (0.33 each), reflecting the analytical judgement that housing conditions, economic vitality, and service provision constitute structurally equivalent pillars of territorial attractiveness:

$$AI_i = \left(IHab_i \times IEcon_i \times ISIG_i\right)^{(1/3)} \quad (4)$$

Table 2. Dimensions and variables of the Attractiveness Index (AI).

Sub-index	Variables used in the Sub-index	Key data sources
Housing (IHab)	(7) H_aloj: *Dwellings per inhabitant. (housing supply)*; H_vaga: *Vacancy rate of conventional dwellings (%)*; H_prec: *Median housing price per* m^2 *(€)*; H_qual: *Share of dwellings with adequate habitability conditions (%)*; H_idade: *Average building age (years)*; H_sec: *Share of secondary residences (%)*; H_tur: *Growth rate of tourist accommodation* (%)	INE Census 2021; INE Local Housing Prices 2024; Tourism Open Data 2025
Economy (IEcon)	(5) E_VAB: *Gross value added per capita (€ per inhabitant)*; E_IPC: *Purchasing power index (%)*; E_des: *Unemployment rate (%)*; E_qual: *Average years of schooling (human capital)*; E_emp: *Employment change (%), 2011–2021*	INE SCIE 2021; INE Purchasing Power 2021; IEFP 2021
Services & Access. (ISIG)	(10) (**Education**): A_esc: *Travel time to secondary schools (minutes)*; A_ebas: *Travel time to basic schools (minutes)*; A_alunos: *Students per school ratio*; A_varEquip: *Variation in school equipment (%)*; (**Healthcare**): A_hosp: *Travel time to nearest hospital (minutes)*; A_ucs; *Travel time to primary healthcare units (minutes)*; A_farm: *Travel time to pharmacies (minutes)*; A_utentes: *Patients per healthcare unit ratio*; A_med; *Physicians per 1,000 inhab*, (**Basic Goods Accessibility**): A_alim: *Travel time to food retail establishments (minutes)*	DGEEC 2021; DGS records; OSM/DGT road network 2018

The Housing Sub-index ($IHab_i$) captures the structural capacity of territories to support residential fixation through the availability, affordability, quality, and functional use of the housing stock. It integrates both supply-side conditions (e.g., dwellings per inhabitant, vacancy rates) and demand-side pressures (e.g., housing prices, tourism dynamics), while also accounting for physical adequacy (habitability and building age) and competing uses of housing (secondary residences).

Overall, the index reflects the extent to which local housing markets facilitate or constrain long-term population retention and attraction.

It is computed as:

$$\begin{aligned} I_{Hab,i} &= 100 \times (0.15\,H_{aloj,i} + 0.10\,H_{vag,i} + 0.25\,H_{prec,i} + 0.15\,H_{qual,i} + \\ &\quad 0.10\,H_{idade,i} + 0.10\,H_{sec,i} + 0.15\,H_{var_tur,i} \end{aligned} \tag{5}$$

The Economy Sub-Index ($IEcon_i$) captures the structural and dynamic capacity of local economies to generate employment, income, and development opportunities. It combines indicators of productive performance (GVA per capita), household well-being (purchasing power), labour market conditions (unemployment rate), and human capital (education levels), alongside recent employment dynamics.

Together, these dimensions reflect both the underlying economic structure and the short-term capacity of territories to sustain and attract population through viable economic prospects.

It is given by:

$$I_{Econ,i} = 100 \times \left(0.25\,E_{VAB,i} + 0.20\,E_{IPC,i} + 0.25\,E_{Des,i} + 0.20\,E_{Qual,i} + 0.10\,E_{VarEmp,i}\right) \tag{6}$$

The General Interest Services and Accessibility Sub-index ($ISIG_i$) captures the extent to which territories ensure effective access to essential services and daily-life infrastructure. It integrates spatial accessibility (travel times), service provision capacity (e.g., healthcare supply, school density), and recent service dynamics (evolution of education, healthcare, and retail networks). The index reflects how service availability and proximity shape quality of life and condition residential decision-making, particularly in low-density contexts where distance and service decline are critical constraints.

It is computed as:

$$I_{\mathrm{SIG},i} = 100 \times \begin{pmatrix} 0.10\,A_{\mathrm{Esc},i} + 0.15\,A_{\mathrm{Ebas},i} + 0.05\,A_{\mathrm{Alunos},i} + 0.05\,A_{\mathrm{VarEquip},i} \\ +0.05\,A_{\mathrm{Hosp},i} + 0.10\,A_{\mathrm{UCP},i} + 0.05\,A_{\mathrm{Farm},i} + 0.15\,A_{\mathrm{Sutentes},i} + 0.05\,A_{\mathrm{Med},i} + 0.25\,A_{\mathrm{Alim},i} \end{pmatrix} \tag{7}$$

Data were collected from official and publicly available sources, including the Portuguese National Statistics Institute (INE), the Institute for Employment and Vocational Training (IEFP), the Directorate-General for Education and Science Statistics (DGEEC), the Directorate-General of Health (DGS), and spatial datasets derived from OpenStreetMap (OSM) and the Directorate-General for Territory (DGT). The most recent available data were used for each variable, accepting minor temporal inconsistencies to ensure analytical relevance.

All variables were normalised using min–max transformation and adjusted for polarity so that higher values consistently represent more favourable territorial conditions.

Variables originally associated with negative territorial outcomes (e.g., unemployment rates, travel times, vacancy rates) were inverted before aggregation.

4 Results

4.1 Demographic Vulnerability Index

The spatial distribution of the DVI reveals a well-defined pattern across the territory. The highest vulnerability scores are concentrated in the interior of North and Centre of Portugal, extending into significant parts of Alentejo. In these municipalities, accelerated population loss, extreme demographic ageing, persistently low fertility, and negative or stagnant net migration reinforce one another, producing particularly severe conditions of demographic risk. The coastal–interior divide is sharp and persistent, consistent with the long-run structural dynamics described in Sect. 2.

Equally important, however, are the exceptions that the DVI reveals. Not all LDT municipalities follow the dominant pattern, a considerable number record intermediate or even low vulnerability scores, suggesting that demographic trajectories within this category are more varied than aggregate narratives imply. This finding carries a direct methodological implication, i.e., low-density classification cannot be treated as a uniform proxy for demographic risk. This internal heterogeneity within the LDT category makes simple binary classifications analytically inadequate, and points to the need for more granular diagnostic tools (Fig. 1).

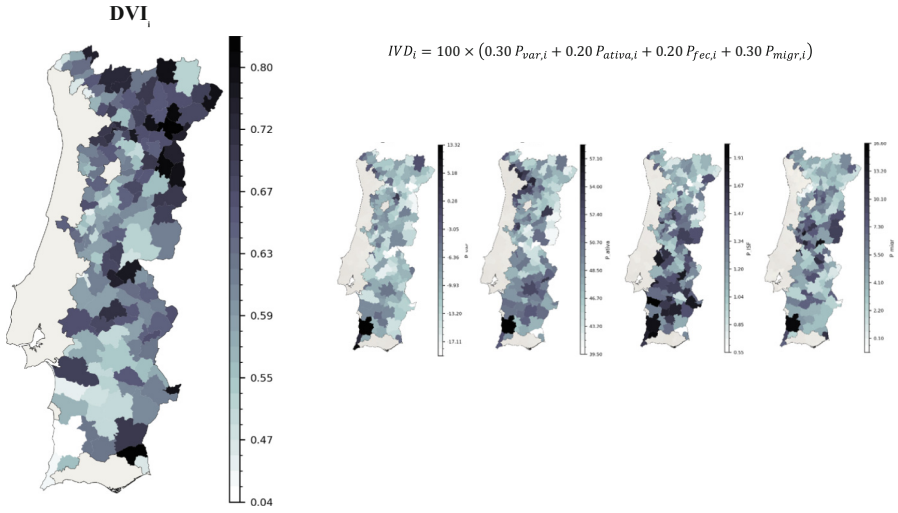


Fig. 1. Demographic Vulnerability Index (DVI) and its underlying variables: population change (P_VAR), working-age population share (P_ENV), total fertility rate (P_ISF), and net migration rate (P_MIGR).

4.2 Attractiveness Sub-Index Results

Turning to housing conditions, the Housing Sub-index (IHab) results reveal a spatial pattern that runs counter to expectations. The highest housing attractiveness scores are not

concentrated along the coast, but rather in the interior North–Centre, the very corridor displaying the greatest demographic vulnerability. This pattern reflects relatively affordable housing markets, adequate building stock quality, and limited pressure from secondary residences and tourism. Coastal municipalities, by contrast, face tourism-driven price inflation and a high concentration of non-primary residential uses, both of which substantially reduce the attractiveness of local housing to permanent residents. What emerges, then, is that housing conditions in many interior LDTs represent a latent asset for population retention, though one that cannot operate in isolation from improvements in economic conditions and service provision (Fig. 2).

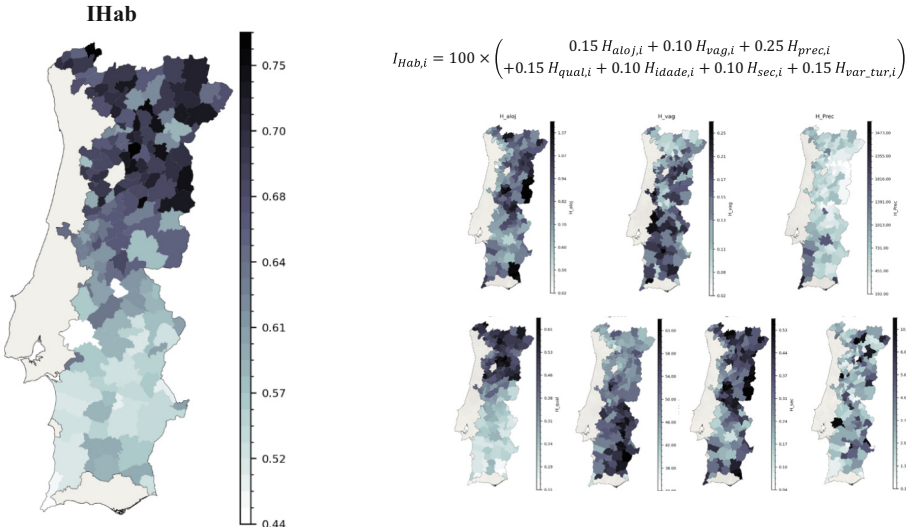


Fig. 2. Housing Sub-index (IHab) and its underlying variables: dwellings per inhabitant (H_ALOJ), vacancy rate (H_VAG), median housing price (H_PREC), habitability conditions (H_QUAL), building age (H_IDADE), secondary residences (H_SEC), and tourist accommodation growth (H_VAR_TUR).

The Economy Sub-index (IEcon) reveals a more heterogeneous territorial pattern. The municipalities with the strongest economic scores combine high productive output (measured by GVA per capita), lower unemployment levels, a more educated workforce, and positive employment dynamics. However, significant asymmetries are observable: in some cases, high economic performance is tied to a narrow set of sector-specific activities and does not translate into broader labour-market strength or human capital development. This shows that sectoral concentration alone is insufficient to ensure long-term territorial attractiveness. The most economically disadvantaged areas have a few things in common: (i) first, they have low productivity, (ii) second, many people there have limited income levels, (iii) third, there is high unemployment and (iv) finally, there

are fewer and fewer jobs available. All of these factors reflect structural constraints that reduce their capacity to sustain and attract population (Fig. 3).

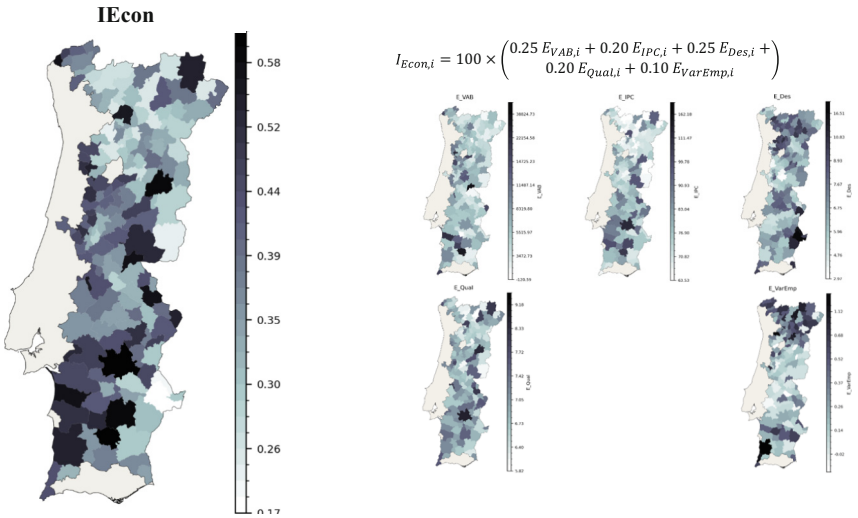


Fig. 3. Economy Sub-index (IEcon) and its underlying variables: gross value added per capita (E_VAB), purchasing power index (E_IPC), unemployment rate (E_DES), mean years of schooling (E_QUAL), and employment change (E_VAREMP).

The Services and Accessibility Sub-index (ISIG) reveals that limited access to healthcare is the most challenging issue for LDT municipalities. Travel times to primary care facilities are long, and there are few medical professionals across much of the interior. Educational accessibility is better in municipalities with viable school networks, though this varies. Access to basic goods, particularly food retail, increasingly depends on the quality of road infrastructure, a finding with direct implications for infrastructure investment priorities in the interior (Fig. 4).

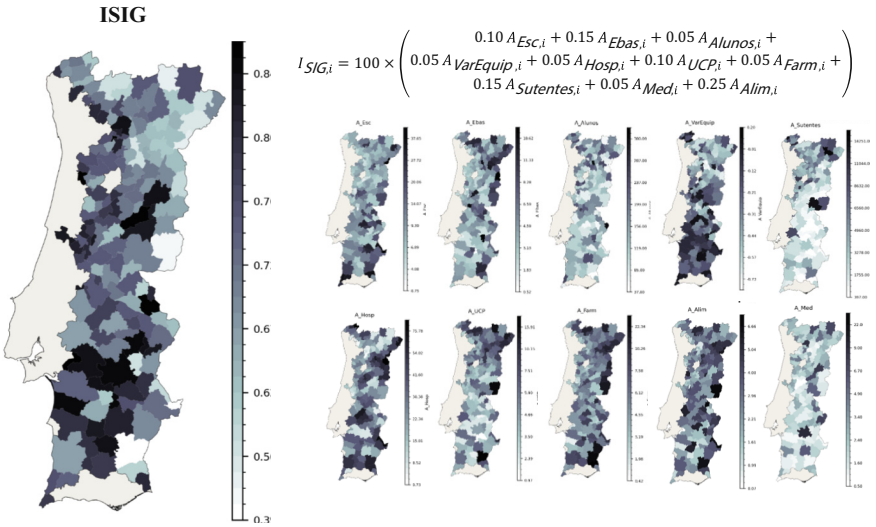


Fig. 4. Services and Accessibility Sub-index (ISIG) and its underlying variables: travel time to primary school (S_EB1), travel time to secondary school (S_EB23), pupils per school (S_ALUNOS), change in school equipment (S_VARESC), travel time to primary healthcare (S_CSP), general practitioners per 1,000 inhabitants (S_MED), change in primary care units (S_VARCSP), travel time to pharmacy (S_FARM), travel time to supermarket (S_SUPER), and change in retail units (S_VARCOM).

4.3 Integrated DVI-AI Typology

Mapping municipalities on a DVI–AI cartesian plane produces a four-quadrant territorial typology with direct policy relevance. Table 2 characterises the four emergent profiles. The largest share of LDT municipalities falls in the Critical (Double Jeopardy) quadrant, combining high demographic vulnerability with low attractiveness. The Fragile / Latent Potential quadrant reveals a meaningful subset of demographically fragile territories that retain significant assets in housing, economy, or services, creating opportunities for targeted activation policies (Fig. 5).

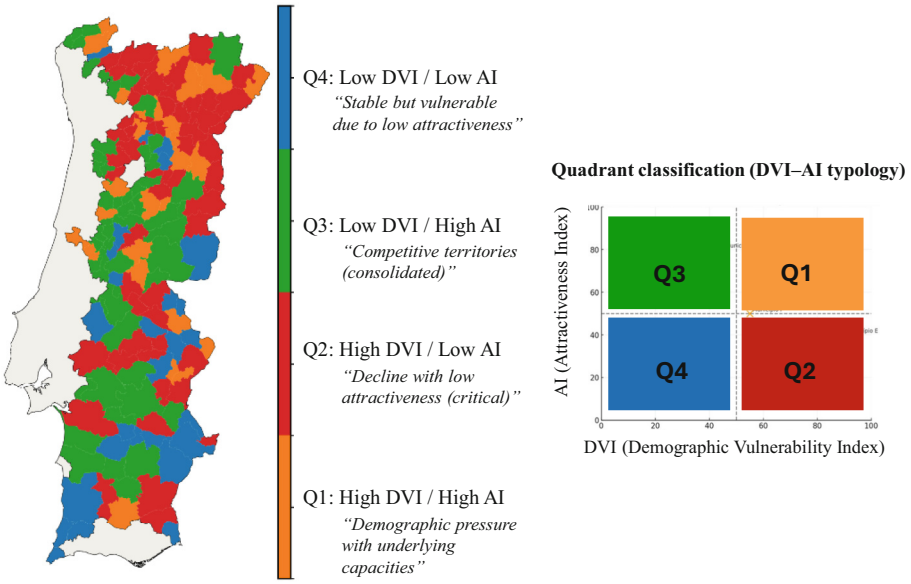


Fig. 5. DVI–AI territorial typology across Portuguese low-density municipalities, showing four quadrants (Q1–Q4) based on the intersection of demographic vulnerability and territorial attractiveness.

The DVI–AI intersection defines four territorial profiles with distinct policy implications: critical areas (high vulnerability, low attractiveness) requiring managed shrinkage strategies; latent potential territories (high–high) suited for activation policies; resilient systems (low vulnerability, high attractiveness) acting as development nodes; and stable but low-attractiveness areas (low–low), where service maintenance and monitoring are key.

5 Discussion

The empirical results support the central analytical claim of this paper, i.e., demographic vulnerability and territorial attractiveness are distinct dimensions that cannot be reduced to one another, yet they interact in ways that matter substantially for territorial diagnosis and policy design. Rather than aligning along a single continuum from decline to vitality, LDT municipalities exhibit a range of configurations, combining varying degrees of structural fragility and latent development potential. This heterogeneity is masked when analytical frameworks rely on single indicators or treat low-density classification as synonymous with uniform decline.

The paper's methodological contribution is the integrated operationalisation of both dimensions within a coherent composite index system. The use of weighted arithmetic means at the sub-index level allows for partial substitutability among closely related indicators within the same domain. The weighted geometric mean applied across dimensions introduces a non-compensatory logic: a territory with severely deficient economic conditions cannot compensate for that weakness by performing well on housing or service

provision. This design produces diagnostics that are more demanding and more policy-realistic than purely additive alternatives. The resulting four-quadrant typology (DVI–AI matrix) translates the analytical output directly into differentiated policy logics, ranging from adaptive management in highly fragile contexts to activation strategies in territories with residual structural assets.

A finding of particular policy significance is that high demographic vulnerability does not reliably coincide with low territorial attractiveness. The concentration of favourable housing conditions in some of the most demographically fragile municipalities points to underexploited assets that conventional analyses, focused predominantly on decline indicators, tend to overlook. This reinforces the case for moving beyond binary typologies and towards more nuanced, multidimensional diagnostics in both research and policy practice.

6 Conclusions

This paper has developed and applied an integrated composite index framework for the joint assessment of demographic vulnerability and territorial attractiveness in low-density territories. Applied to Portugal's 165 officially classified LDT municipalities, the analysis produces a structurally differentiated territorial map. Demographic fragility is most pronounced in the North–Centre–Alentejo interior corridor, yet the distribution of territorial attractiveness conditions varies considerably within this space, far more than aggregate narratives of interior decline would suggest.

The most significant finding is the systematic disconnection between vulnerability and attractiveness. Some of the most demographically fragile municipalities have favourable housing conditions, which are overlooked by standard analytical frameworks. This discrepancy challenges deterministic accounts of territorial decline, creating analytical space for more selective, opportunity-oriented policy approaches.

The results also make clear that effective responses to demographic decline in low-density territories cannot rely on uniform growth-oriented strategies. What is needed are territorially differentiated approaches: ones that protect quality of life under conditions of demographic contraction, mobilise endogenous assets where they exist, and selectively invest in territorial anchor points with demonstrated capacity for broader diffusion effects. In this context, adaptation, rather than reversal, is the more realistic and analytically defensible policy paradigm.

By integrating multiple dimensions into a transparent, replicable composite index system, the framework developed here provides a practical diagnostic tool for identifying territorial imbalances and informing evidence-based regional policy. Its methodological design is sufficiently flexible to extend beyond the Portuguese context, offering potential for comparative application across low-density regions elsewhere in Europe facing analogous structural challenges.

Acknowledgments. The authors acknowledge the support of PLANAPP – Centre for Planning and Evaluation of Public Policies – within the framework of the Science4Policy 2025 programme. We also thank all members of the REVITA project team for their valuable contributions, collaboration, and ongoing commitment throughout the development of this work.

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

References

1. Recano, J.: Despoblamiento y despoblación: Dos conceptos complementarios. Boletín de la Asociación de Geógrafos Españoles **97** (2023)
2. Coleman, D., Rowthorn, R.: Who is afraid of population decline? a critical examination of its consequences. Popul. Dev. Rev. **37**(Suppl. 1), 217–248 (2011)
3. Dumont, G.-F.: Depopulation as a global structural trend. Population et Sociétés (forthcoming, 2025)
4. Gomes, M.C., Pimentel, P.: Demographic dynamics in Portugal: structure, trends and territorial differentiation. Análise Social **57**(245), 840–870 (2022)
5. Hospers, G.-J.: Policy responses to urban shrinkage: from growth thinking to civic engagement. Eur. Plan. Stud. **22**(7), 1507–1523 (2014)
6. Pallagst, K.: Shrink Smart: the Research Contribution of the European Research Network. Cardiff University, Cardiff (2009)
7. Wiechmann, T., Pallagst, K.: Urban shrinkage in Germany and the USA: a comparison of transformation patterns and local strategies. Int. J. Urban Reg. Res. **36**(2), 261–280 (2012)
8. Syssner, J.: Quietly Shrinking Cities: Swedish Planning in Contexts of Urban and Rural Population Decline. Policy Press, Bristol (2022)
9. Hollander, J.: Sunburnt Cities: the Great Recession. Depopulation and Urban Planning in the American Sunbelt. Routledge, London (2011)
10. Béal, V., Fol, S., Epstein, R., Rousseau, M.: French shrinking cities: urban policies and politics of denial. Town Plan. Rev. **90**(1), 1–21 (2019)
11. Bretagnolle, V., et al.: Combining demographic and economic indicators for territorial diagnosis. J. Reg. Sci. **59**(4), 701–724 (2019)
12. Instituto Nacional de, Estatística: Population and Housing Census 2021. INE, Lisbon (2021)

Identifying Optimal Locations for Shared e-Scooters Stations in Lexington, Kentucky

Eirini Stavropoulou[1](✉), Dionysios Tzamakos[1], Lambros Mitropoulos[1], and Nikiforos Stamatiadis[2]

[1] School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Zografou Campus, 15780 Athens, Greece
{eirini_stavropoulou,dionysios_tzamakos,lmit}@mail.ntua.gr
[2] Department of Civil Engineering, University of Kentucky, Lexington, KY, USA
nick.stamatiadis@uky.edu

Abstract. In several urban areas, the scarcity of shared micromobility stations to serve first- and last-mile trips coupled with inadequate connections to public transport, limit access to a significant portion of the city's residents. While some dockless schemes are available, these primarily cater to the needs of those in the city center, leaving other areas unserved. This study develops a spatially driven approach to identify suitable areas and optimal locations for deploying first- and last- mile shared micromobility services. The proposed method is applied to the city of Lexington, Kentucky, United States. Initially, a spatial analysis is performed to identify suitable areas for locating shared e-scooter stations. Subsequently, a network analysis is conducted to identify a limited set of optimal station locations within these areas. The outcome of this study is a practical, two-step methodology that does not require large amounts of input data and can be used by micromobility companies and transportation planners when planning and launching new shared micromobility systems.

Keywords: E-scooters · Micromobility · Data-driven · Suitability Index · Optimal Location · Network Analysis

1 Introduction

Over the last years, e-scooter sharing services have expanded rapidly in many cities around the world, primarily to serve first- and last-mile trips. Since 2017, cities across the United States (US) have introduced shared e-scooter sharing programs that often co-exist with docked bikeshare services [1]. This rapid uptake is reflected in usage statistics. In 2018, e-scooters accounted for the highest proportion of micromobility trips in the US, with 38.5 million recorded trips [2, 3]. By 2024, shared e-scooter ridership continued its rapid growth, reaching 85 million trips across North America, a 38% increase from the previous year [4].

Since the use of shared micromobility has been spreading recently all around the world, there are still many open issues in terms of suitable infrastructure, demand analysis

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 470–486, 2027.
https://doi.org/10.1007/978-3-032-30494-0_30

and safety evaluations [5]. The wide spatiotemporal availability of these services, made possible by mobile applications that easily match real time demand and supply, favors the use of micromobility in combination with other modes of transport, such as fixed and demand-responsive transit. Integrating micromobility with public transport could also enlarge its catchment area, increase its accessibility, and reduce congestion phenomena and pollution, thus making cities more livable [6].

Regarding micromobility infrastructure, the positioning of shared e-scooter stations with respect to suitably aligned criteria is a crucial element in the effectiveness and sustainability of micromobility networks [7]. In this context, the aim of this study is to introduce a methodology for planning first- and last-mile micromobility services by identifying the optimal locations for shared e-scooter stations. The selection process is guided by five spatial criteria: (i) population density, (ii) land-use characteristics, (iii) proximity to bike lanes, (iv) proximity to bust stations, and (v) proximity to major Points of Interest (POIs). According to García-Palomares et al. [8] e-scooter station locations must be related to the proximity of population, activities and the public transport network. The proposed methodology is applied in Lexington – Fayette Urban County, Kentucky (KY) in the US. The results of this study can serve as an effective tool to help urban planners design an equitable and profitable shared docked e-scooter system, since it attempts to pinpoint locations where there is a need for e-scooters, thus maximizing their usage.

2 Literature Review

Shared e-scooter services are a subset of shared micromobility systems. Shared mobility refers to "the shared use of a vehicle, bicycle or other low-speed travel mode" [9] and is often described as a product-service system, as it provides access to a mobility asset (e.g., an e-scooter) without requiring ownership [10]. In the literature, micromobility is defined as the use of micro-vehicles; vehicles with a mass not exceeding 350 kg and a maximum design speed of 45 km/h." [11].

The popularity of micromobility-shared systems has been rising in cities all over the world due to a number of advantages [7]. Especially e-scooters, these include flexibility, quick access to and from public transport stops, energy efficiency, reduced noise levels, minimal parking space requirements and no exhaust pollution [12]. In particular, shared e-scooters offer a convenient solution for short-distance trips and first- and last-mile connections, potentially reducing dependence on private automobiles and alleviating traffic congestion in dense urban environments [13]. As a result, micromobility has been widely discussed as a promising component of more sustainable and efficient urban transport systems [14].

The potential environmental and mobility benefits of micromobility, however, are closely tied to the transport modes it replaces [15]. The findings of previous studies regarding the transport modes that are mainly substituted by micromobility services are contrasting. Studies conducted in European countries indicate that micromobility services often replace walking and, to a lesser extent, public transport trips [16, 17]. Therefore, in many cases the net emission balance is negative and the benefits of micromobility for the environment are questionable. In contrast, mode substitution patterns in

the United States differ, with studies reporting that micromobility services have managed to replace a substantial number of private car trips [13, 18, 19]. Despite the contrasting results, the vast majority of the studies acknowledges that e-scooters' characteristics are well suitable for the first-last mile of any trip chain [20]. Consequently, several studies have examined the integration of micromobility with public transport services and the overall potential of micromobility to enhance multimodality and accessibility [21, 22].

From an infrastructure planning perspective, the spatial configuration of shared micromobility systems plays a critical role in their performance. Most shared micromobility systems in the United States operate as docked systems, in which stations are located in places designated by local authorities [7, 23]. Compared to dockless or free-floating systems, docked e-scooter systems are more oriented toward regular and commuter-oriented travel patterns [23]. Each of these systems requires specific stations where the vehicles will be securely locked. Decisions regarding the initial placement and size of these stations, as well as strategies for expanding the system by adding new stations or reorganizing existing ones, are critical [7]. Prior studies emphasize that establishing a successful shared service network requires careful consideration of station locations relative to POIs, public transportation, and overall user accessibility [24, 25]. Empirical evidence from Washington D.C., for example, indicates that higher population density and greater concentrations of POIs are associated with increased e-scooter usage [26]. Consequently, a quantitative site suitability analysis for shared e-scooter services is essential.

As a result, the identification of optimal locations for shared micromobility infrastructure in urban environments has become an active area of research. Existing studies address both dockless and docked micromobility systems, often with the objective of maximizing system usage while minimizing operational costs and maintaining high service quality [27]. Methodologically, the literature identifies that the use of spatial models for the prediction of the optimal location of facilities has the ability to overcome the weaknesses of linear regression models, as they better account for spatial autocorrelation (i.e., the correlation between two different points in space) and non-stationarity [28].

A portion of the literature focuses on data-intensive, demand-driven approaches. For instance, Sandoval et al. [29] proposed a data-driven method to identify parking locations for shared dockless e-scooters based on the potential demand by clustering historical trip data. Likewise, Zakhem and Smith-Colin [30] used trip clustering method to identify parking zones for dockless e-scooters. Colovic et al. [31] developed a Multi-objective Micromobility Maximal Coverage Parking Location (M-MCPL) model that returns the optimal shared e-scooter parking locations. This model is based on the following objective functions: (i) the maximization of the population coverage; (ii) the maximization of multimodal accessibility coverage (i.e., bus, railway, and metro modes); (iii) the maximization of the attraction coverage considering the most relevant points of interest for each corresponding zone in large urban areas. Similarly, Deveci et al. [32] used a hybrid fuzzy Multi – Criteria Decision Making (MCDM) model considering 12 criteria, concerning users, public authority, service operator, and urban livability. Finally, Ayfantopoulou et al. [33] conducted a data-driven analysis using the K-Means algorithm to

select suitable locations for docked scooter-sharing service stations, based on trip data, road geometric characteristics, and station capacity.

A smaller body of research employs GIS-based analyses to identify optimal facility locations. Roslan and Naharudin [7], for example, developed a GIS-based MCDM approach to optimally positioning rental stations of docked e-scooter shared services in respect to prospective consistent criteria. They used eight criteria, including population density and proximity to significant locations, to perform a suitability analysis in Shah Alam city in Malaysia. Likewise, Kabak et al. [34] deployed a similar approach to locate bike-sharing station sites in Karsiyaka, Izmir, Turkey. While these studies highlight the value of spatially explicit, criteria-driven methods, they often depend on extensive datasets or context-specific assumptions that may constrain their transferability.

Overall, the literature demonstrates that while shared e-scooter systems offer meaningful mobility and sustainability benefits, their effectiveness depends heavily on the spatial configuration of the supporting infrastructure. Prior research highlights the importance of station placement for ensuring accessibility, encouraging multimodal integration, and maximizing system usage. A wide range of methodological approaches has been explored, from demand-driven clustering techniques to multi-criteria decision-making models. Yet most studies require extensive and context-specific datasets. In contrast, comparatively few studies develop GIS-based frameworks tailored specifically to shared micromobility systems. This gap underscores the need for practical, transferable methods that can guide early-stage planning in data-constrained environments. Responding to this need, the present study proposes a GIS-based approach that integrates area-level suitability analysis with network-based location optimization, using a limited but meaningful set of criteria derived from the existing literature to support planners and operators in the strategic deployment of docked e-scooter infrastructure.

3 Methodology

3.1 Study Area and Data Collection

The proposed methodology is applied to the case study of Lexington – Fayette Urban County, KY. Lexington is the second largest city in the state, with a population of nearly 320,000 residents [35]. Most people in Lexington commute to work with car, pickup truck, or van (86.1%), while only 1.8% use public transportation in 2020 according to the US Census Bureau [35]. This strong reliance on private vehicles highlights the need for alternative mobility options that can enhance first- and last-mile connectivity, particularly in conjunction with public transit.

Public transportation in Lexington is provided by Lextran, which operates a fleet of 65 buses along a network of 22 bus routes [36]. The bus network follows a hub-and-spoke configuration, with routes primarily oriented towards downtown Lexington as a central hub and extending outward in a radial pattern. While this configuration supports downtown access, it limits cross-town connectivity and reduces transit accessibility for areas located farther from the central core.

At present, shared micromobility services in Lexington consist exclusively of dockless e-scooters. The absence of docked e-scooter and bike-share stations, combined with inadequate integration between micromobility services and the existing public transport network, constrains the potential of e-scooters to effectively serve first- and last-mile trips. This limit access to a significant portion of the city's residents.

Data used in this study refer to the Lexington – Fayette Urban County and were collected from various publicly available sources. First, the census block groups boundaries coupled with population totals from the 2020 census are acquired from the ESRI Data Hub [37]. Additional spatial datasets utilized in this study include land cover data [38], bicycle and pedestrian facilities (BKPD) [39], and the location of bus stops and POIs, both sourced from OpenStreetMap (OSM) [40]. Population data refer to 2022 measurements, land cover data are from 2021, the BKPD was produced in 2019 and OSM data reflect the most recent updates available in 2024. The use of these widely accessible datasets aligns with the study's goal of developing a transferable methodology that does not rely on extensive data sources.

All spatial analyses were conducted using ArcGIS Pro version 3.1.0. Three main extensions were employed: Spatial Analyst, Geoprocessing Tools and Network Analyst. Data preprocessing steps included clipping all datasets to the study area boundary and ensuring a consistent projected coordinate system across layers. Figure 1 presents the spatial distribution of population and population density at the census block group level, highlighting a clear core-periphery pattern in Lexington–Fayette Urban County, with higher population counts and especially higher population densities concentrated in the urban core and surrounding inner neighborhoods, while outer and peripheral block groups exhibit lower densities despite sometimes containing larger total populations. Figure 2 illustrates land cover patterns across the study area, emphasizing the strong contrast between the densely developed urban core – dominated by medium- and high-intensity developed land – and the surrounding peripheral areas, which are largely characterized by agricultural land uses (hay/pasture and cultivated crops) interspersed with forested patches and limited low-intensity development.

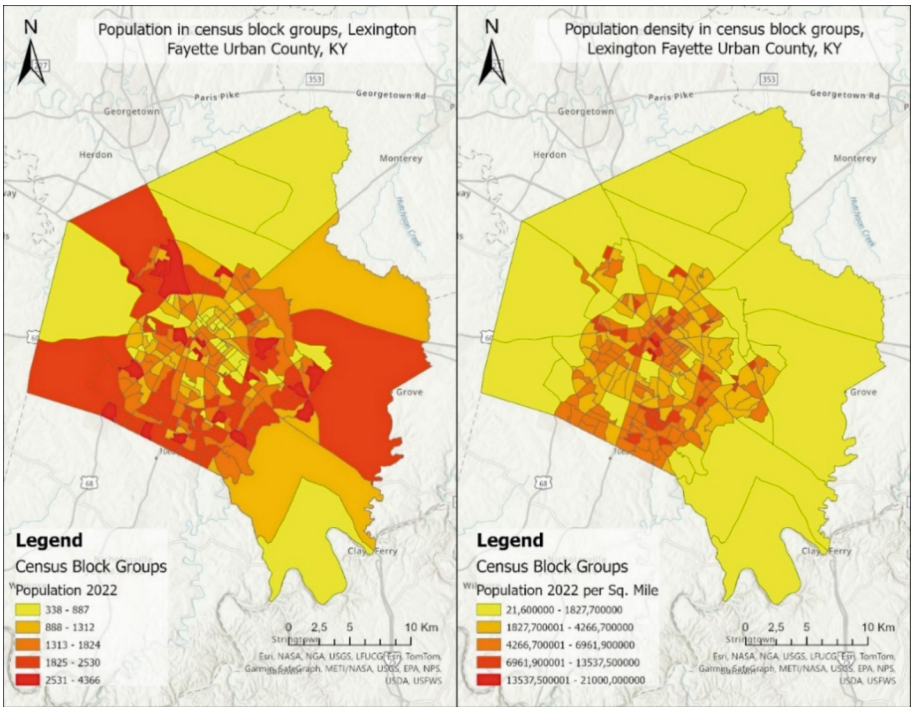


Fig. 1. Population and Population Density in census block groups, Lexington Fayette Urban County, KY (Data source: [37], Maps created by the authors)

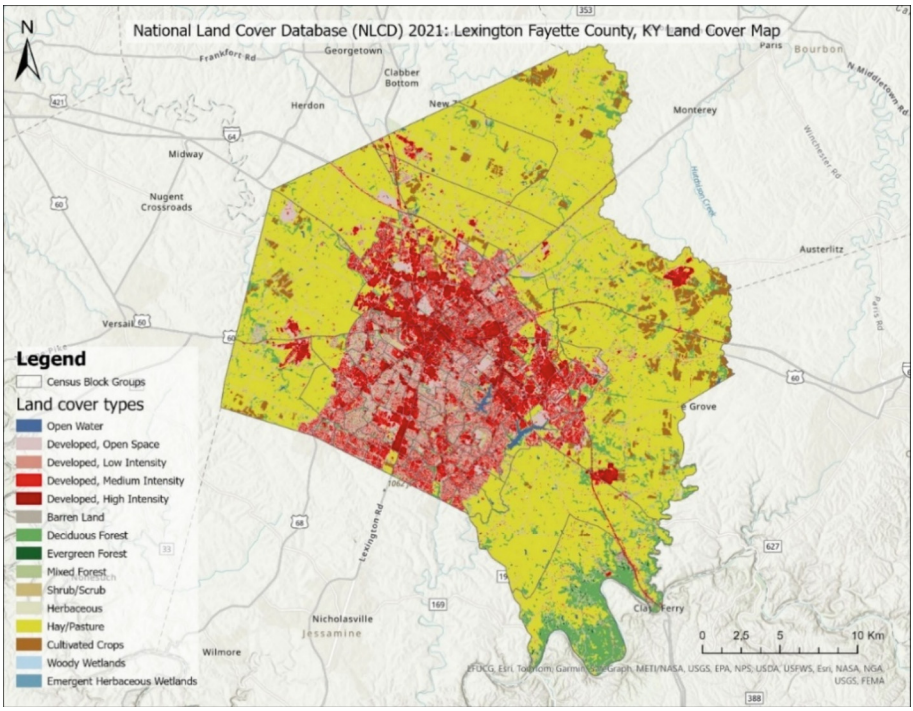


Fig. 2. Land cover in Lexington Fayette Urban County, KY (Data source: [38], Map created by the authors)

3.2 Suitability Analysis

The first objective of this study is to identify suitable areas within Lexington where shared e-scooter stations can be located, with the aim of maximizing potential usage. This is achieved through a suitability analysis, a spatial decision-support process in which multiple criteria are evaluated and combined to identify locations that best satisfy predefined objectives. The selection and weighting of criteria are critical steps in such analyses, as different factors may contribute unequally to overall suitability.

In the literature, several techniques have been proposed to determine the relative importance (weights) of suitability criteria, including the Analytic Hierarchy Process (AHP) and other multi-criteria decision-making approaches [41]. In this study, however, all selected criteria are assigned equal weights. This choice is made deliberately to maintain methodological simplicity, and to avoid introducing subjective bias in the absence of stakeholder-specific preferences. The focus of the analysis is therefore placed on the spatial interaction of the criteria rather than on weight calibration.

The suitability analysis is modelled by using a Weighted Overlay Analysis (WOA) in which several rasters are overlayed by using a common measurement scale and each one is weighted according to its importance. Based on the literature review, five criteria are selected to spatially define the suitable areas for shared e-scooter station deployment: (i) population density, (ii) land use, (iii) proximity to bike lanes, (iv) proximity to bus stations, and (v) proximity to POIs. For population density, areas with higher densities are

assumed to be more suitable, reflecting higher potential demand. For land use, the highest suitability score is assigned to "Developed, High Intensity" areas, as these locations typically exhibit concentrated activities and trip generation.

The remaining three criteria are proximity-based and are calculated using Euclidean distance analysis. This method computes the distance from each raster cell to the nearest relevant feature (bike lanes, bus stations, or POIs). In all cases, shorter distances are considered more favorable, as closer proximity enhances accessibility and the likelihood of e-scooter usage.

The overall suitability modeling framework is illustrated in Fig. 3 and consists of several sequential steps: rasterization of each vector layer, calculation of distance measurements where applicable, and reclassification of each criterion. The reclassification of each feature involves the rating of each cell from score 0 (not suitable) to score 10 (most suitable) in six classes. The five criteria, reclassified and rated, are presented in Fig. 4.

For the land-use criterion, certain land cover classes are treated as exclusionary variables. Specifically, land cover types deemed unsuitable for station placement (e.g., water bodies or undeveloped natural areas) are assigned a "NoData" value during reclassification. This operation removes these cells from subsequent calculations, ensuring that they do not contribute to the final suitability index.

Following reclassification, the five standardized, unitless raster layers are combined using the WOA to produce a composite suitability index for the study area. The resulting suitability index for the area, shown in Fig. 5, represents the spatial distribution of suitability for shared e-scooter station deployment across Lexington and serves as the basis for the subsequent network-based location optimization.

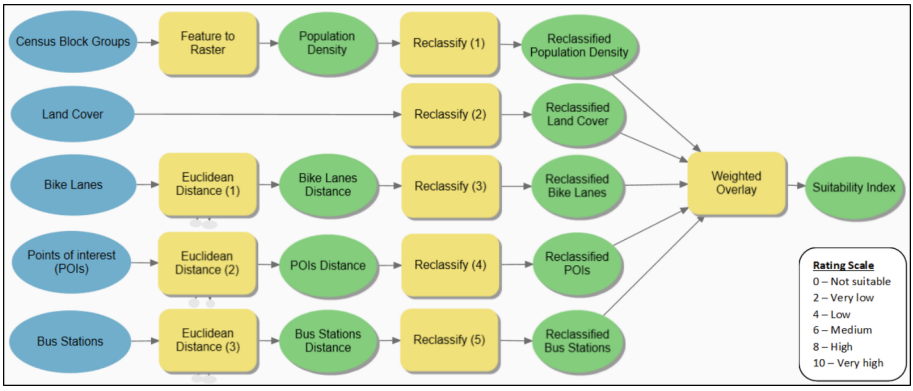


Fig. 3. Suitability modeling workflow

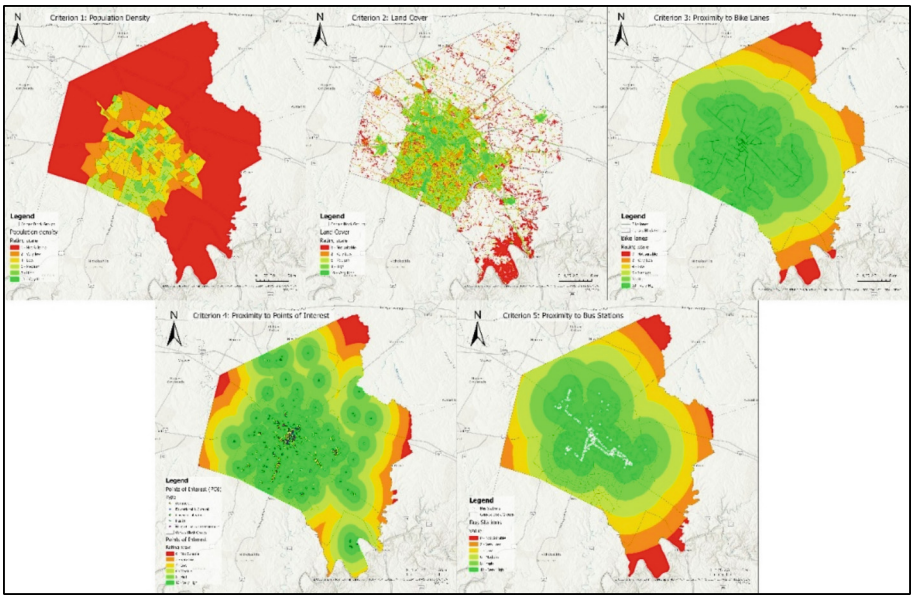


Fig. 4. The five (5) criteria for the suitability analysis (after reclassification)

3.3 Network Analysis

The second objective of this study is to identify the optimal point locations for the shared e-scooter stations within Lexington Fayette Urban County, KY. For that purpose, a network-based location-allocation analysis is employed. Location-allocation models are used in transportation and facility planning to determine the best locations for a predefined number of facilities while allocating demand in a way that optimizes a specific objective function [42].

The analysis builds directly on the results of the suitability assessment. The first step involves extracting the areas identified as highly suitable from the suitability analysis, having a score of 10. Subsequently, this raster layer is converted into polygon features, and then into point features, which serve as potential or candidate facility locations for shared e-scooter stations. These candidate points represent locations that satisfy all suitability criteria and are therefore eligible for final selection.

Demand points are represented by the centroids of census block groups, weighted by population density. This reflects the assumption that areas with higher population concentrations generate greater potential demand for shared e-scooter services. The walking time is adopted as the impedance (cost) metric and the maximum walking time of 10 min is set as the service cutoff. This value is consistent with commonly accepted thresholds for first- and last-mile access to transit and micromobility services [43].

This location-allocation model is configured to select ten facility locations, employing the "Maximize Coverage" approach. This number was chosen to represent a feasible initial deployment scenario for a docked e-scooter system in Lexington. It should be noted that the proposed methodology is scalable, and the number of facilities can be adjusted to reflect alternative planning scenarios. The maximal coverage location problem, originally formulated by Church & Velle [44], seeks to locate a fixed number of facilities so as to maximize the total demand covered within a specified service distance or time [45, 46]. In this context, coverage is defined as the population-weighted demand that can reach a selected e-scooter station within the 10-min walking threshold. The result of this model is presented in Fig. 7, in results section.

4 Results

The suitability index for identifying optimal areas for shared e-scooter stations within Lexington Fayette Urban County, KY is presented in Fig. 5. This index emerged from the WOA with equal weights assigned to each of the five criteria: population density, land use, proximity to bicycle infrastructure, proximity to bus stops, and proximity to POIs. The resulting map illustrates the spatial distribution of suitability areas across the county, with darker green areas indicating higher suitability scores.

High-suitability areas are primarily concentrated in downtown Lexington and in the vicinity of the University of Kentucky. This spatial pattern can be attributed to the convergence of several favorable conditions in these areas, including high population density, mixed and high-intensity land uses, dense concentrations of POIs, and strong proximity to both public transport stops and bicycle infrastructure.

Areas with a suitability score of 0 represent areas that do not meet the minimum criteria for station placement and are therefore correctly identified as not suitable. Regarding a red-colored area in the west of the downtown core between suitable areas, it is deemed unsuitable because the population density in this area is extremely low as it is an industrial area.

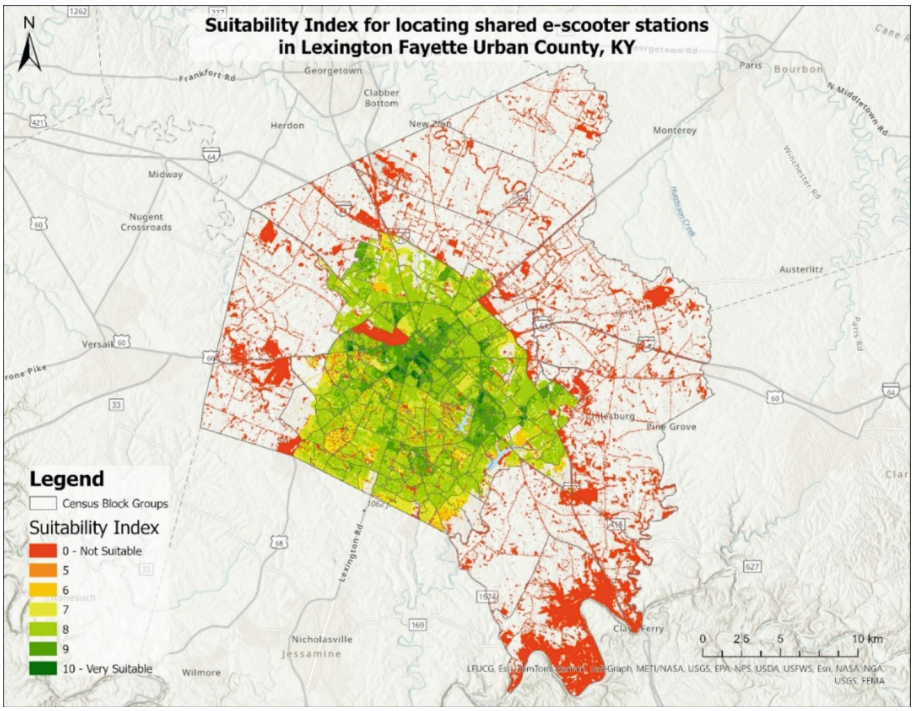


Fig. 5. Suitability index for locating shared e-scooter stations in Lexington Fayette Urban County, KY (WOA with equal weights)

To better understand the model results, a suitability index with unequal weights (Table 1) is also calculated and presented in Fig. 6. In this case, population density and proximity to bike lanes criteria have higher weights. As can be observed from the map in Fig. 6, the unequal weights lead to slightly different results. Specifically, the most suitable areas are fewer and there are more areas rated with the score of 9 out of 10.

Table 1. Criteria weights value

Criteria	Weights
Population Density	0.30
Land cover	0.10
Proximity to bike lanes	0.30
Proximity to POIs	0.10
Proximity to bus stations	0.20
Sum	1

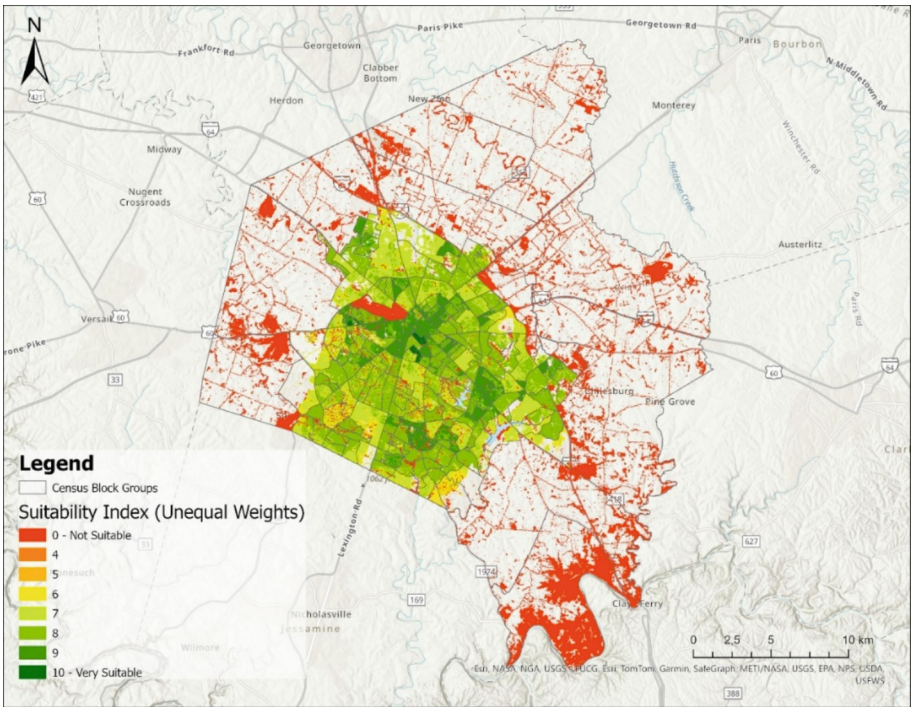


Fig. 6. Suitability index for locating shared e-scooter stations in Lexington Fayette Urban County, KY (WOA with unequal weights)

Figure 7 shows the ten optimal locations, in terms of points, within the suitable areas that are the chosen facilities from the potential ones after the location-allocation model. These locations are optimal, since they are accessible from a high proportion of commuters (maximization of coverage). It is noted that suitable areas are the areas with score ten in suitability analysis with equal weights.

The ten points are indeed in areas with high expected demand, as indicated by high population density, proximity to major points of interest, and accessibility to public transportation and bicycle infrastructure. Notably, one of the points is located within the University of Kentucky, a significant hub of demand in Lexington, with two additional points located in proximity.

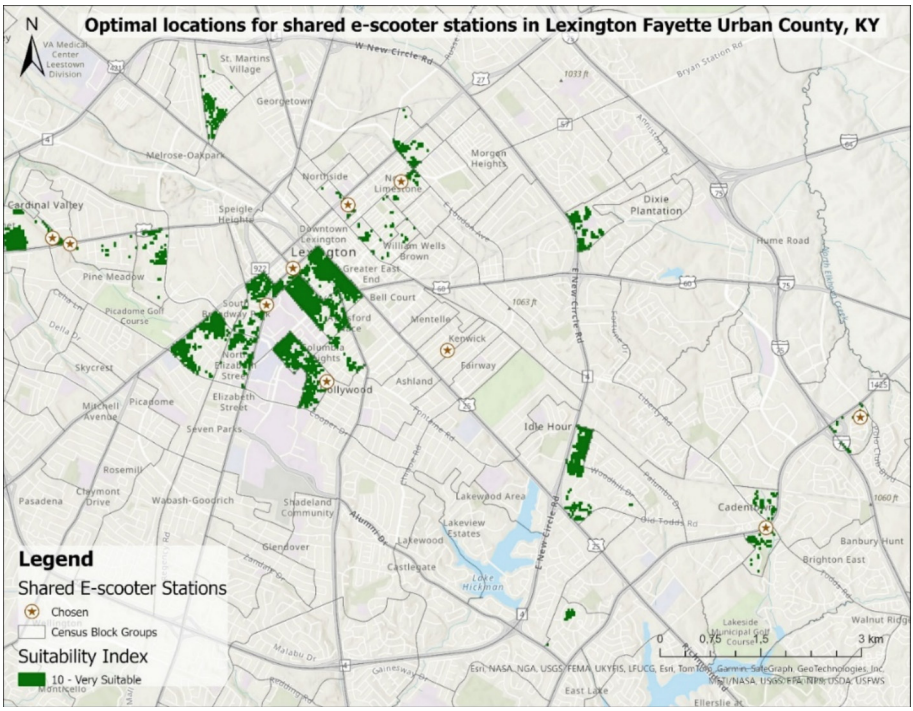


Fig. 7. Optimal locations for shared e-scooter stations in Lexington Fayette Urban County, KY

5 Conclusions

This study presents a comprehensive, two-stage methodology for identifying both suitable areas and optimal point locations for shared e-scooter stations. The methodology is applied to Lexington-Fayette Urban County, KY, a city where docked shared micromobility services are currently absent, and reliance on private vehicles is high. The framework is designed to support planners and micromobility operators in developing data-informed strategies for first- and last-mile mobility solutions.

In the first stage, a WOA is used to generate a suitability index based on five key criteria: population density, land use, proximity to bike lanes, proximity to bus stops, and proximity to POIs. Areas with high suitability scores correspond to locations with high population concentrations, accessible infrastructure, and significant trip-generating activities. This index provides insights into where demand for shared e-scooters is likely to be highest, offering a practical tool for identifying priority areas for system deployment. Users of this methodology have the option to conclude their analysis at this stage. However, for those seeking to identify optimal points rather than areas, the second part of the methodology offers a solution.

The second stage of the methodology applies a location-allocation model to the candidate areas derived from the suitability analysis. Using a maximize coverage approach, the model identifies ten optimal locations for docked e-scooter stations, ensuring that the largest proportion of population-weighted demand is within a 10-min walking distance.

The selection of ten stations represents a feasible initial deployment scenario. Notably, the identified locations correspond to areas with high population density, proximity to major activity hubs such as the University of Kentucky, and convenient access to public transport and bicycle infrastructure.

Although this methodology offers significant utility, several limitations should be acknowledged. The analysis currently relies on a fixed set of five criteria with equal weighting and does not explicitly account for sociodemographic characteristics such as age, gender, or car ownership, which could influence demand patterns. Similarly, detailed travel demand data could further refine the location-allocation results [8]. In addition, the framework combines established methods (weighted overlay analysis and maximize-coverage location-allocation) and does not benchmark its performance against more advanced approaches, such as demand clustering or stakeholder-driven multi-criteria decision-making models. The proposed station locations are also not empirically validated against observed micromobility usage patterns, expert judgement, or alternative parameter settings beyond a limited sensitivity test, as docked systems are not currently implemented in the study area. Furthermore, demand is primarily approximated through population density and accessibility variables, without explicit equity metrics or demographic segmentation.

Future research could expand the set of criteria, incorporate stakeholder preferences for weighting, and conduct broader sensitivity and benchmarking analyses across alternative modeling approaches. The integration of observed or proxy demand data (e.g., from dockless systems or travel surveys) would enable empirical validation of the results, while the inclusion of equity indicators and demographic segmentation could support more inclusive planning. Finally, evaluating alternative numbers of stations through sensitivity analyses would improve the proposed framework by optimizing coverage and operational efficiency.

Overall, this study contributes a scalable, data-light, and transferable framework for planning shared e-scooter facilities in urban environments. It combines geospatial suitability analysis and optimization to provide both planners and mobility operators with a structured approach to identify priority areas and specific locations that maximize service accessibility, connectivity, and potential usage.

Acknowledgments. The author(s) received no financial support for the research.

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

References

1. Meng, S., Brown, A.: How shared e-scooter programs affect docked bikeshare ridership in communities of concern: a tale of two cities. Transportation (2024). https://doi.org/10.1007/s11116-024-10473-w
2. Huo, J., Yang, H., Li, C., Zheng, R., Yang, L., Wen, Y.: Influence of the built environment on *E*-scooter sharing ridership: a tale of five cities. J. Transp. Geogr. **93**, 103084 (2021). https://doi.org/10.1016/j.jtrangeo.2021.103084

3. Younes, H., Zou, Z., Wu, J., Baiocchi, G.: Comparing the temporal determinants of dockless scooter-share and station-based bike-share in Washington, D.C. Transp. Res. Part A Policy Pract. **134**, 308–320 (2020). https://doi.org/10.1016/j.tra.2020.02.021
4. NABSA: NABSA's 6th Shared Micromobility Report Shows Ridership Sets New Record of 225 Million Trips – eBikes Int'l. https://ebikes-international.com/nabsas-6th-shared-micromobility-report-shows-ridership-sets-new-record-of-225-million-trips/. Accessed 14 Jan 2026
5. Fazio, M., Giuffrida, N., Le Pira, M., Inturri, G., Ignaccolo, M.: Planning suitable transport networks for e-scooters to foster micromobility spreading. Sustainability. **13**, 11422 (2021). https://doi.org/10.3390/su132011422
6. Oeschger, G., Carroll, P., Caulfield, B.: Micromobility and public transport integration: the current state of knowledge. Transp. Res. Part D Trans. Environ. **89**, 102628 (2020). https://doi.org/10.1016/j.trd.2020.102628
7. Roslan, A., Naharudin, N.: Identification of e-Scooter Shared (ESS) Stations by using a GIS-based MCDM Approach. IJG 69–78 (2023). https://doi.org/10.52939/ijg.v19i5.2663
8. García-Palomares, J.C., Gutiérrez, J., Latorre, M.: Optimizing the location of stations in bike-sharing programs: a GIS approach. Appl. Geogr. **35**, 235–246 (2012). https://doi.org/10.1016/j.apgeog.2012.07.002
9. Cohen, A., Shaheen, S.: Planning for Shared Mobility. American Planning Association, Chicago (2018)
10. Ruhrort, L.: Reassessing the role of shared mobility services in a transport transition: can they contribute the rise of an alternative socio-technical regime of mobility? Sustainability **12**, 8253 (2020). https://doi.org/10.3390/su12198253
11. ITF: Safe Micromobility. International Transport Forum (2020)
12. Yanocha, D., Allan, M.: The Electric Assist: Leveraging E-bikes and E-scooters for More Livable Cities - Institute for Transportation and Development Policy. ITDP (2019)
13. Wang, K., Qian, X., Fitch, D.T., Lee, Y., Malik, J., Circella, G.: What travel modes do shared e-scooters displace? a review of recent research findings. Transp. Rev. **43**, 5–31 (2023). https://doi.org/10.1080/01441647.2021.2015639
14. Comi, A., Polimeni, A.: Assessing potential sustainability benefits of micromobility: a new data driven approach. Eur. Transp. Res. Rev. **16**, 19 (2024). https://doi.org/10.1186/s12544-024-00640-6
15. Stavropoulou, E., Mitropoulos, L., Tzouras, P.G., Karolemeas, C., Kepaptsoglou, K.: An evaluation of agent-based models for simulating e-scooter sharing services in urban areas. In: Nathanail, E.G., Gavanas, N., Adamos, G. (eds.) Smart Energy for Smart Transport. CSUM 2022. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-23721-8_79
16. Laa, B., Leth, U.: Survey of E-scooter users in Vienna: who they are and how they ride. J. Transp. Geogr. **89**, 102874 (2020). https://doi.org/10.1016/j.jtrangeo.2020.102874
17. Weschke, J., Oostendorp, R., Hardinghaus, M.: Mode shift, motivational reasons, and impact on emissions of shared e-scooter usage. Transp. Res. Part D Transp. Environ. **112**, 103468 (2022). https://doi.org/10.1016/j.trd.2022.103468
18. Fearnley, N.: Factors affecting e-scooter mode substitution. Findings (2022). https://doi.org/10.32866/001c.36514
19. Mitropoulos, L., Stavropoulou, E., Tzouras, P., Karolemeas, C., Kepaptsoglou, K.: E-scooter micromobility systems: review of attributes and impacts. Transp. Res. Interdisciplinary Perspect. **21**, 100888 (2023). https://doi.org/10.1016/j.trip.2023.100888
20. Tzouras, P.G., et al.: Agent-based models for simulating e-scooter sharing services: a review and a qualitative assessment. Int. J. Transp. Sci. Technol. **12**, 71–85 (2023). https://doi.org/10.1016/j.ijtst.2022.02.001

21. Javadiansr, M., Davatgari, A., Rahimi, E., Mohammadi, M., Mohammadian, A. (Kouros), Auld, J.: Coupling shared E-scooters and public transit: a spatial and temporal analysis. Transp. Lett. 1–18 (2023). https://doi.org/10.1080/19427867.2023.2227447
22. Yan, X., Zhao, X., Broaddus, A., Johnson, J., Srinivasan, S.: Evaluating shared e-scooters' potential to enhance public transit and reduce driving. Transp. Res. Part D Transp. Environ. **117**, 103640 (2023). https://doi.org/10.1016/j.trd.2023.103640
23. Nawaro, Ł: E-scooters: competition with shared bicycles and relationship to public transport. Int. J. Urban Sustain. Dev. **13**, 614–630 (2021). https://doi.org/10.1080/19463138.2021.1981336
24. Yang, L., et al.: Space-time demand cube for spatial-temporal coverage optimization model of shared bicycle system: a study using big bike GPS data. J. Transp. Geogr. **88**, 102861 (2020). https://doi.org/10.1016/j.jtrangeo.2020.102861
25. Mitropoulos, L., Nikiforiadis, A., Fotiou, A.M., Kortsari, A., Ayfantopoulou, G.: An accessibility-driven method for planning first and last mile micromobility services. Transp. Res. Proc. **90**, 266–273 (2025). https://doi.org/10.1016/j.trpro.2025.06.067
26. Hawa, L., Cui, B., Sun, L., El-Geneidy, A.: Scoot over: determinants of shared electric scooter presence in Washington D.C. Case Stud. Transp. Policy **9**, 418–430 (2021). https://doi.org/10.1016/j.cstp.2021.01.003
27. Fathabad, A.M., Li, X., Cheng, J., Wu, Y.-J.: Data-Driven Optimization for E-Scooter System Design. Transportation Research and Education Center (TREC) (2022). https://doi.org/10.15760/trec.274
28. Efthymiou, D., Antoniou, C., Tyrinopoulos, Y.: Spatially aware model for optimal site selection: method and application in a Greek mobility center. Transp. Res. Rec. **2276**, 146–155 (2012). https://doi.org/10.3141/2276-18
29. Sandoval, R., et al.: Data driven methods for effective micromobility parking. Transp. Res. Interdisciplinary Perspect. **10**, 100368 (2021). https://doi.org/10.1016/j.trip.2021.100368
30. Zakhem, M., Smith-Colin, J.: Micromobility implementation challenges and opportunities: analysis of e-scooter parking and high-use corridors. Transp. Res. Part D Transp. Environ. **101**, 103082 (2021). https://doi.org/10.1016/j.trd.2021.103082
31. Colovic, A., Prencipe, L.P., Giuffrida, N., Ottomanelli, M.: A multi-objective model to design shared e-kick scooters parking spaces in large urban areas. J. Transp. Geogr. **116**, 103823 (2024). https://doi.org/10.1016/j.jtrangeo.2024.103823
32. Deveci, M., Gokasar, I., Pamucar, D., Chen, Y., Coffman, D.: Sustainable E-scooter parking operation in urban areas using fuzzy Dombi based RAFSI model. Sustain. Cities Soc. **91**, 104426 (2023). https://doi.org/10.1016/j.scs.2023.104426
33. Ayfantopoulou, G., Salanova Grau, J.M., Maleas, Z., Siomos, A.: Micro-mobility user pattern analysis and station location in Thessaloniki. Sustainability **14**, 6715 (2022). https://doi.org/10.3390/su14116715
34. Kabak, M., Erbaş, M., Çetinkaya, C., Özceylan, E.: A GIS-based MCDM approach for the evaluation of bike-share stations. J. Clean. Prod. **201**, 49–60 (2018). https://doi.org/10.1016/j.jclepro.2018.08.033
35. US Census Bureau: Census.gov. https://www.census.gov/en.html. Accessed 17 Apr 2024
36. Lextran. https://lextran.com/. Accessed 02 Apr 2024
37. ESRI hub: USA Census Block Group Boundaries. https://hub.arcgis.com/content/1c924a53319a491ab43d5cb1d55d8561/about. Accessed 18 Apr 2024
38. MRLC: Multi-Resolution Land Characteristics (MRLC) Consortium. https://www.mrlc.gov/data?f%5B0%5D=category%3ALand%20Cover. Accessed 18 Apr 2024
39. KYTC: DataMart. https://datamart.kytc.ky.gov/. Accessed 18 Apr 2024
40. Geofabrik: Open Street Map Data for Kentucky. https://download.geofabrik.de/north-america/us/kentucky.html. Accessed 18 Apr 2024

41. Hassan, I., et al: Weighted overlay based on land suitability analysis of argiculture land land in AJK using GIS and AHP. Pakistan J. Argicult. Sci. (2020)
42. Dehdari Ebrahimi, Z., Momenitabar, M., Nasri, A.A., Mattson, J.: Using a GIS-based spatial approach to determine the optimal locations of bikeshare stations: the case of Washington D.C. Transp. Policy **127**, 48–60 (2022). https://doi.org/10.1016/j.tranpol.2022.08.008
43. McQueen, M., Clifton, K.J.: Assessing the perception of E-scooters as a practical and equitable first-mile/last-mile solution. Transp. Res. Part A Policy Pract. **165**, 395–418 (2022). https://doi.org/10.1016/j.tra.2022.09.021
44. Church, R., Velle, C.R.: The maximal covering location problem. Pap. Reg. Sci. **32**, 101–118 (1974). https://doi.org/10.1111/j.1435-5597.1974.tb00902.x
45. Momenitabar, M., Dehdari Ebrahimi, Z., Arani, M., Mattson, J., Ghasemi, P.: Designing a sustainable closed-loop supply chain network considering lateral resupply and backup suppliers using fuzzy inference system. Environ. Dev. Sustain. (2022). https://doi.org/10.1007/s10668-022-02332-4
46. Safaei, S., Ghasemi, P., Goodarzian, F., Momenitabar, M.: Designing a new multi-echelon multi-period closed-loop supply chain network by forecasting demand using time series model: a genetic algorithm. Environ. Sci. Pollut. Res. **29**, 79754–79768 (2022). https://doi.org/10.1007/s11356-022-19341-5

Green and Grey Transition Data Analysis for Developing Synergies in the RECs Context

Lorenza Mola[1], Gianmarco Lorenti[2], Giorgia Malavasi[3], Lisa Panero[3], Amina Pereno[3], Maurizio Repetto[2], Diana Rolando[3(✉)], and Martina Spinelli[3]

[1] Dipartimento di Giurisprudenza Campus Luigi Einaudi, Università di Torino, Lungo, Dora Siena 100, Torino, Italy
[2] Department of Energy (DENERG) "Galileo Ferraris", Politecnico di Torino, Corso Duca Degli Abruzzi, 24, 10129 Torino, Italy
[3] Department of Architecture and Design (DAD), Politecnico di Torino, Viale P.A. Mattioli 39, 10125 Torino, Italy
diana.rolando@polito.it

Abstract. Addressing climate change and socio-territorial inequalities requires integrated approaches that connect green and grey transitions within local development processes. In this context, Renewable Energy Communities (RECs) offer a strategic opportunity, particularly in fragile and mountain areas. This study aims to investigate how integrated data analysis can support the identification of synergies between energy systems, service provision, and local demand. The research adopts a spatially integrated approach based on the combination of socio-demographic, territorial, and energy-related datasets within a Spatial Information System (SIT). Through a collaborative and interdisciplinary process, the framework enables the interpretation and matching of multiple data layers. This paper investigates the interaction between demographic ageing, socio-spatial vulnerability, and energy transition processes in the Nort-Western Canavase, a mountain area in Piedmont (Italy) characterised by depopulation and limited accessibility to welfare and energy infrastructures. Analytical mapping and accessibility assessments are employed to identify critical areas and potential synergies. The analysis highlights the added value of integrating green (energy-related) and grey (demographic ageing and service contraction-related) dimensions. It reveals mismatches between energy supply and local demand, as well as opportunities to enhance service provision through energy initiatives. Areas characterised by high socio-territorial vulnerability emerge as priority contexts for targeted actions. The study demonstrates the potential of integrated data analysis to support decision-making in the RECs context. By fostering interdisciplinary collaboration and evidence-based insights, the proposed framework contributes to more inclusive, place-based, and synergic transition strategies.

Keywords: Homecare Services · Energy Vulnerability · Active Ageing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 487–502, 2027.
https://doi.org/10.1007/978-3-032-30494-0_31

1 Introduction

The transition toward low-carbon energy systems represents one of the central challenges of contemporary territorial development [1]. In recent years, particular attention has been devoted to the role of Renewable Energy Communities (RECs), increasingly framed as key instruments for fostering decentralised energy production, enhancing local resilience, and supporting the green transition. In this sense, the "green transition" is understood as a territorial process of decarbonization and energy system transformation [2] through the deployment of renewable energy infrastructure and sustainable energy policies [3]. However, in inner and rural areas, this transition unfolds within contexts characterised by structural demographic and socio-economic vulnerabilities, raising critical questions about its feasibility, inclusiveness, and long-term sustainability [4].

At the same time, many European inner regions are experiencing a parallel and equally significant process of demographic ageing and service contraction, here defined as the grey transition. In this paper, the term "grey transition" refers to the demographic transformation associated with population ageing, depopulation [5], and increasing socio-spatial vulnerability [6, 7], particularly affecting marginal and peripheral territories [8]. In such contexts, socio-assistential services are not only essential infrastructures for care provision, but also key elements shaping territorial cohesion and quality of life [9]. While the "green transition" refers to the structural transformation towards low-carbon and renewable energy systems [10], the concept of "grey transition" has been increasingly used to describe demographic ageing, shrinking labour forces, and rising dependency ratios affecting European societies [8].

Despite their coexistence, green and grey transitions are often addressed through separate policy and analytical frameworks. Energy transition strategies tend to focus on technological and infrastructural aspects, while socio-assistential systems are typically analysed within welfare and public service domains. This separation overlooks the fact that both processes are deeply interconnected at the territorial level. Demographic ageing directly influences energy demand patterns, service accessibility needs, and infrastructural pressures, while energy systems, particularly when organised through Solidarity and Renewable Energy Communities (CERS), can potentially support local welfare systems, reduce energy poverty, foster social cohesion and contribute to the provision of services in fragile contexts [11, 12].

Within this framework, the SustAge project (led by the Law Department of the University of Turin, [13]) investigates the relationship between green and grey transition, with a specific focus on the integration between energy communities and home-based socio-assistential services in rural and mountain areas. The project adopts a place-based and interdisciplinary perspective, aiming to bridge the gap between energy, social, and territorial dimensions.

The central research problem lies in the disconnection between sectoral approaches to energy transition and the socio-territorial conditions in which they are implemented. In particular, existing strategies often overlook the role of demographic dynamics and socio-assistential systems in shaping both energy needs and implementation capacity. The study is therefore guided by the hypothesis that green and grey transitions are structurally interdependent, especially in fragile rural areas. However, previous studies rarely integrate demographic vulnerability, socio-assistential accessibility, and renewable energy potential within a unified GIS-based analytical framework.

This paper focuses on the first preliminary results of the SustAge project, to demonstrate how socio-economic and energy data analysis serve as a foundational methodological step for identifying constraints and opportunities in a specific territorial context, and for supporting the development of integrated strategies of local governance that address the interdependencies between green and grey transition.

The research is grounded in a place-based approach, which recognises territorial specificity as a fundamental resource for designing effective and context-sensitive interventions. This perspective is complemented by an interdisciplinary framework integrating contributions from legal studies, systemic design, economic evaluation, and electrical engineering.

The North-Western Canavese area (Piedmont, Italy) was selected as a case study for its emblematic combination of demographic fragility, territorial fragmentation, and energy potential.

The main objectives are:

- to develop a comprehensive territorial analysis integrating multiple data layers;
- to identify criticalities and latent potentials within the study area;
- to explore the interactions between energy systems and socio-assistential structures;
- to provide a robust analytical basis for future phases of the project.

The paper is structured as follows. Section 2 presents the methodological approach, outlining the integrative and interdisciplinary framework adopted to combine demographic, socio-economic, and energy-related data within a spatial analytical environment. Section 3 introduces the case study of the North-Western Canavese, highlighting its territorial, demographic, and infrastructural characteristics. Section 4 discusses the main results of the analysis, focusing on interactions between grey and green transitions and identifying patterns of synergies and mismatches. Finally, Sect. 5 provides the conclusions, summarising the key findings and outlining future research directions.

2 Methodological Approach

The study adopts a multi-disciplinary framework, based on the integration of data analyses on green and grey transition, assuming the concept of holistic territorial diagnosis [14]. This approach enables the combination of heterogeneous datasets from different sources into a coherent analytical framework [15], allowing for the identification of interdependencies across spatial, social, and technical dimensions. Rather than treating each analytical dimension separately, this approach emphasises their systemic integration [16], in line with the complexity of the case study. The proposed framework should be interpreted as an exploratory GIS-based analytical approach combining quantitative spatial indicators with a qualitative interpretation of territorial vulnerability.

The methodological approach spatially integrates heterogeneous socio-demographic, infrastructural, and energy-related datasets in order to identify territorial patterns and criticalities.

Within this framework, demographic dynamics are interpreted not merely as descriptive variables, but as drivers of territorial transformation [17], directly influencing demand for care services, mobility requirements, and energy consumption patterns. At

the same time, accessibility is conceptualised as a multi-dimensional concept, related to energy and technical systems, infrastructure provision, and service distribution [18]. It is therefore understood as a key determinant of territorial cohesion and social inclusion, particularly for vulnerable populations [19]. The integration of economic, social, and technical dimensions allows for a more comprehensive understanding of how structural ageing processes interact with energy energy systems and renewable potential, potentially mitigating territorial inequalities [20].

In this perspective, the grey transition is not only a demographic process linked to ageing, but also a spatial phenomenon [21], reshaping both the composition of the population and the accessibility of services across the territory.

Figure 1 illustrates the methodological framework adopted to integrate green and grey transition. The approach combines demographic and energy datasets through a spatial data integration layer, enabling the identification of synergies and mismatches [22] between service demand, accessibility conditions, and energy potential.

The methodological framework is structured around the convergence of two distinct but interrelated domains: the grey transition, associated with demographic dynamics and socio-assistential systems, and the green transition, related to energy systems and renewable potential. The grey transition is analysed using demographic and social data [20], including population structure, ageing patterns, and the distribution of service demand and supply, while the green transition is examined through energy-related data [23, 24], such as renewable energy potential, existing infrastructures, and energy system configurations.

These two domains are brought together within a data elaboration and integration phase, in which heterogeneous datasets are combined through spatial analysis techniques. This phase relies on the overlay and cross-referencing of demographic, service, and energy data, enabling a systemic interpretation of the territory and revealing underlying spatial patterns.

Building on this integrated dataset, the framework advances through a collaborative [25] and interdisciplinary interpretation and matching process, engaging different disciplines. In particular, the data integration enables the assessment of the alignment between energy supply and local demand, the interaction between service provision and energy potential, and the identification of areas characterised by high socio-territorial vulnerability [26, 27].This interpretative step makes it possible to detect both synergies and mismatches between systems that are typically analysed separately.

In this context, a SWOT analysis is employed as a synthetic evaluative tool to systematically identify strengths, weaknesses, opportunities, and threats emerging from the interaction between socio-demographic and energy-related dimensions.

The resulting outputs consist of a spatial classification of territorial patterns, distinguishing areas of socio-territorial fragility, areas with high renewable energy potential, areas with significant service or energy demand, as well as critical mismatch areas where demand is not adequately met by existing supply, and potential synergy areas where energy production capacity overlaps with social and service hubs. Overall, the approach enables a shift from sectoral analyses toward a relational and integrative understanding of territorial systems [14, 28], providing a robust analytical basis for the development of place-based strategies that simultaneously address environmental and social challenges.

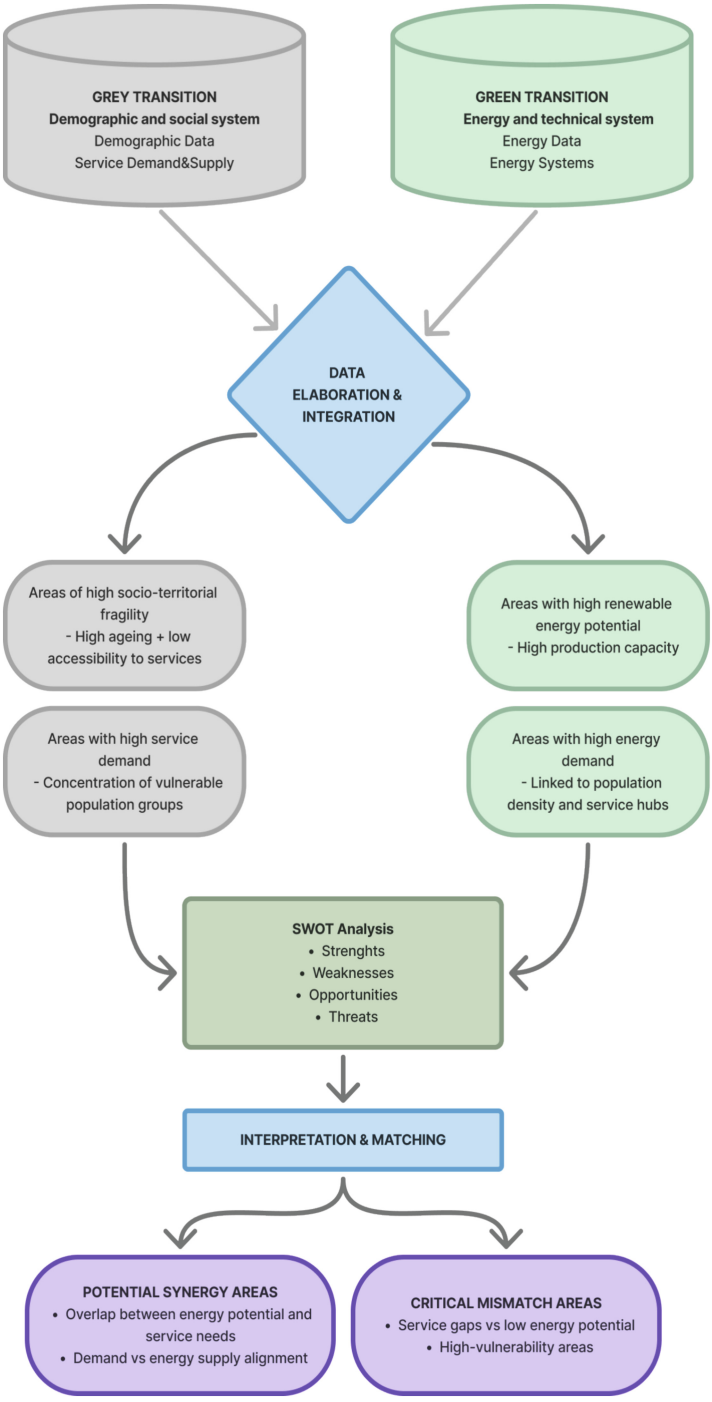


Fig. 1. Methodological framework adopted in the present research (Authors' elaboration).

3 Case Study

The North-Western Canavese represents a typical rural context, where geographical, demographic, and infrastructural conditions interact to produce complex territorial dynamics. The territory is predominantly characterised by mountain and semi-mountain areas, with numerous small municipalities distributed across valleys and upland zones. This morphology contributes to a dispersed settlement pattern, which, combined with limited transport connectivity, generates significant challenges in terms of accessibility to services and infrastructures.

From a demographic perspective, the area exhibits a pronounced ageing trend, with a high proportion of elderly population and a progressive decline in younger cohorts. This dynamic directly affects the demand for healthcare and social services, placing increasing pressure on local welfare systems. In this context, the presence of the CISS38 socio-assistential consortium [29] plays a crucial role, acting as a central coordinating body for the provision and management of social services across the territory (Fig. 2). Despite this institutional structure, the spatial distribution of services remains uneven, particularly affecting more remote and mountainous municipalities, where accessibility constraints amplify conditions of socio-territorial fragility [30].

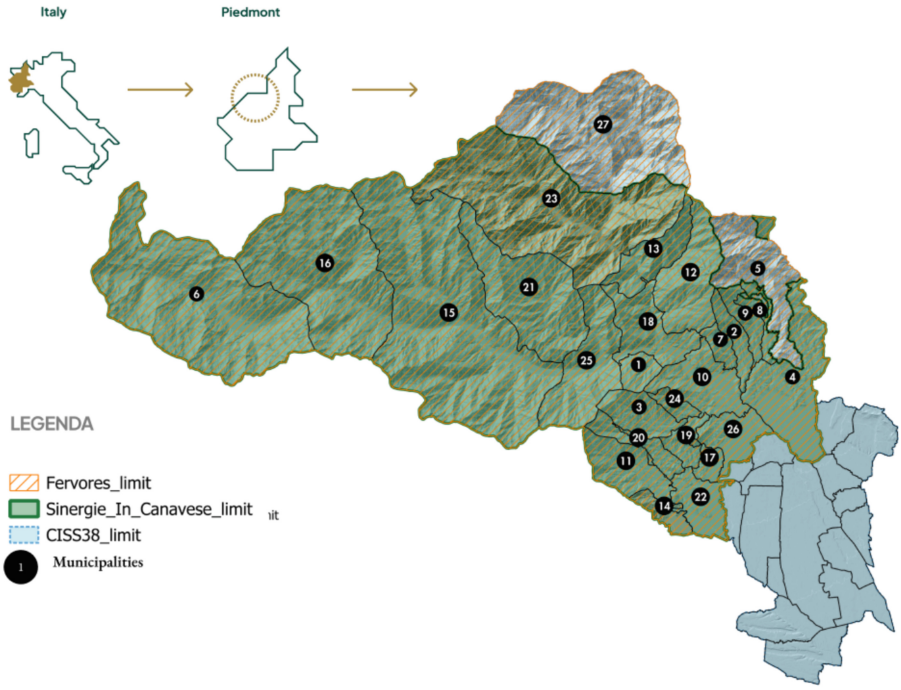


Fig. 2. North-Western Canavese area (Authors' elaboration on Geoportale Torino - Banca Dati Territoriale di Riferimento degli Enti (BDTRE), 2025).

At the same time, the territory presents relevant opportunities in the field of energy transition. It is important to highlight the presence of the Green Community *Sinergie in Canavese* [31], which promotes cooperation among municipalities to advance sustainable management, enhance energy efficiency, support green tourism, and counteract depopulation, and the Fervores CERS [32], which reflects a growing capacity to activate local resources and experiment with decentralised energy models. These initiatives highlight the potential of energy systems not only as technical infrastructures, but also as instruments for fostering territorial cohesion and supporting local communities.

Overall, the coexistence of ageing population, infrastructural constraints, and significative energy production capacity makes the North-Western Canavese a particularly relevant context for exploring the integration between green and grey transition. The interaction between demographic dynamics, service provision, and energy systems suggests the need for place-based and interdisciplinary approaches capable of addressing environmental and social challenges in a coordinated manner.

4 Results

The integration of data analyses on green and grey transition did not primarily result in a static representation of territorial conditions, but rather in a process of knowledge construction based [28] on the interaction between different disciplinary perspectives and data layers. A key outcome of this phase was the visualisation of demographic, socio-assistential, infrastructural, and energy-related data within a shared analytical environment, enabling a shift from sectoral interpretations to a systemic reading of the territory.

In this context, the collaboration between disciplines played a central role. Each domain contributed distinct datasets, interpretative frameworks, and analytical priorities, which were progressively aligned through processes of data elaboration, visualisation, and mapping. Data visualisation [33], in particular, emerged as a critical methodological tool, allowing complex and heterogeneous information to be translated into readable and comparable spatial representations. Through the development of a Spatial Information System (SIT), it became possible to highlight relationships that would remain otherwise implicit, such as the overlap between demographic vulnerability, service accessibility, and energy potential.

The data elaboration and integration phase was therefore not limited to the technical combination of datasets, but involved an interpretative process in which mapping practices supported visual outputs. Maps and visual representations functioned as boundary objects, enabling different disciplinary perspectives to converge around shared evidence and facilitating the identification of relevant analytical dimensions. Building on this SIT, the subsequent phase of interpretation and matching enabled the identification of relationships between green and grey transition, moving beyond their conventional separation. The analysis revealed how demographic ageing, service demand, and accessibility conditions interact with energy production potential, infrastructural configurations, and consumption patterns. Rather than treating these domains independently, the framework enabled explicit matching of demand and supply across social and energy systems (Fig. 3).

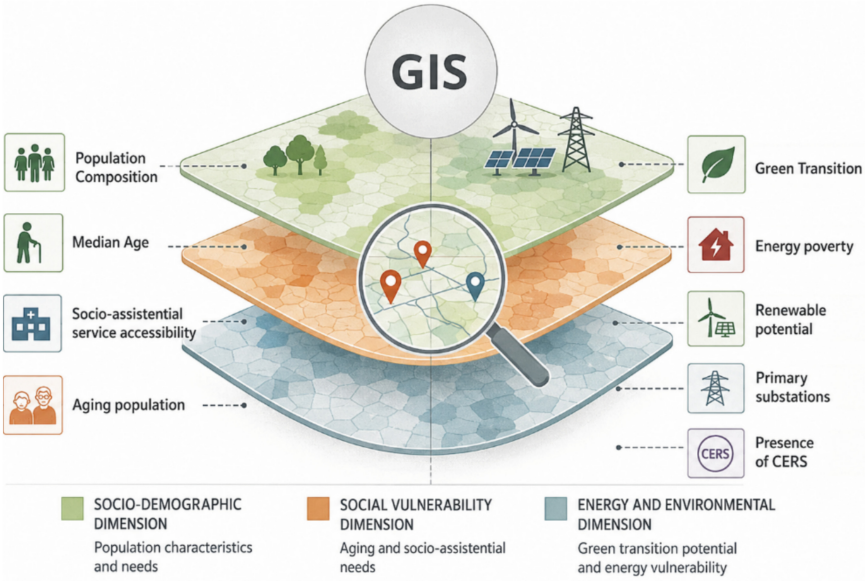


Fig. 3. Conceptual framework of the Spatial Information System (SIT) integrating grey and green transition dimensions. (Authors' elaboration, generated with AI support, OpenAI DALL·E).

The results of this matching process consisted of the identification of different types of territorial conditions, defined by the degree of alignment or misalignment between the two transitions. In particular, the analysis highlighted situations in which high levels of socio-demographic vulnerability correspond to limited service accessibility and to the fragmentation of the social fabric, reflected in instances of isolation among older populations, often associated with economic hardship and infrastructural limitations in accessing energy ('energy poverty' [34]). It also identified contexts in which energy production potential coincides with significant social demand, particularly in areas where spaces for social and healthcare interaction are already present, or where new hubs may be required to support community- and home-based care services.

At the same time, the analysis identified areas in which the overlap between energy potential and social needs suggests the possibility of integrated interventions, particularly through decentralised and community-based networks. In this sense, the results provided not only a diagnostic reading of the territory, but also an analytical foundation for linking energy transition strategies with welfare and service provision systems [23].

Overall, the results demonstrated that the added value of the proposed methodological approach lies in its capacity to connect heterogeneous data through visualisation and interdisciplinary interpretation, enabling the identification of potential synergies and critical mismatch areas between green and grey transition and supporting the development of place-based and integrated strategies.

4.1 Demographic Ageing and Socio-Assistential Services

The analysis of demographic dynamics and socio-assistential services highlighted the structural relevance of the grey transition within the study area. The territory is characterised by a markedly ageing population, with a high proportion of residents over 65 years (28.04%) and a significant imbalance between older and younger groups (309 people aged 65 or over for every 100 young people aged 0 to 14 in 2025). This demographic configuration translates into an increasing demand for healthcare, social assistance, and proximity-based services, particularly in municipalities affected by population decline and settlement dispersion (Fig. 4).

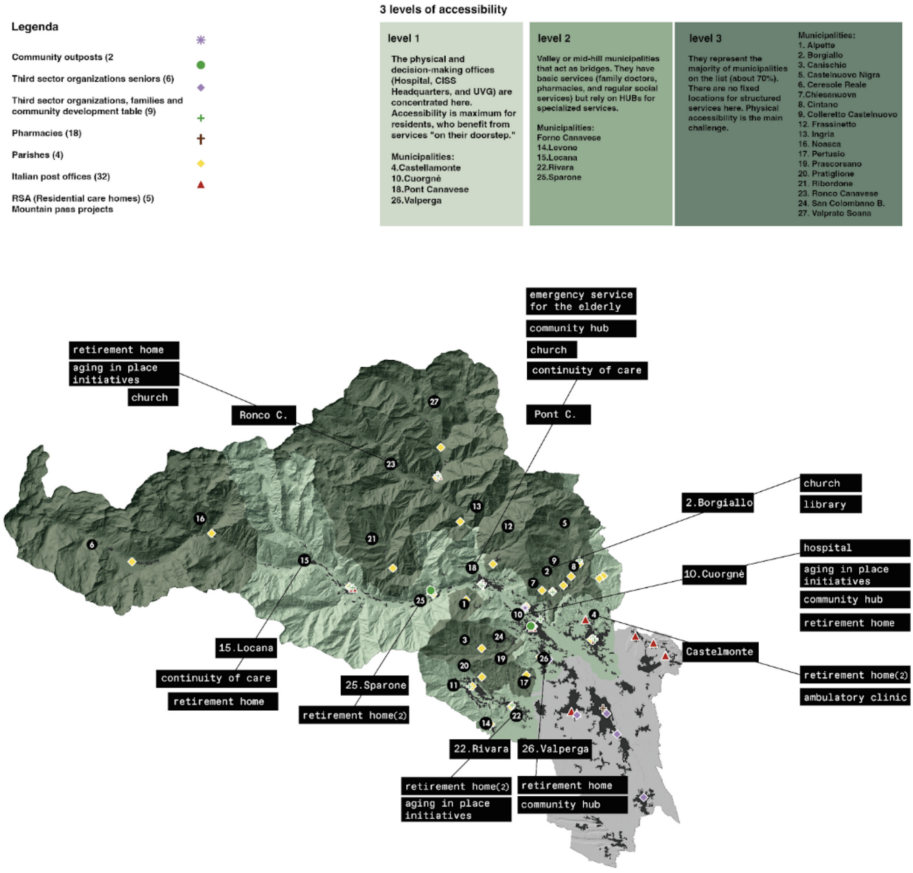


Fig. 4. Socio-Assistential System in the study area (Authors' elaboration). Cartographic sources: Geoportale Piemonte – Banca Dati Territoriale di Riferimento degli Enti (BDTRE). Data sources: CISS38, Progetto "Passi Montanti", and on-desk surveys.

The mapping of socio-assistential services, combined with accessibility analysis, revealed a multi-level and spatially uneven system of provision. Central municipalities (Cuorgnè and Castellamonte) concentrate the main service infrastructures, including

healthcare facilities and administrative functions, while intermediate areas rely on partial and often periodic services. Peripheral and mountain municipalities (Ceresole Reale, Locana and Noasca), on the other hand, are characterised by limited physical accessibility and the absence of structured service provision, relying instead on mobile, home-based, or emergency services.

Within this framework, the presence of the CISS38 socio-assistential consortium (in Cuorgnè) plays a crucial role in coordinating service provision across the territory. However, the analysis showed that institutional coordination alone is not sufficient to overcome territorial constraints related to distance, morphology, and infrastructural limitations. Accessibility emerged as a key mediating factor, shaping the effective capacity of the system to respond to demographic demand.

Through data visualisation and mapping, it was possible to highlight the misalignment between areas of higher demographic vulnerability and the spatial distribution of services, particularly in mountain areas. This confirms that the grey transition is related not only to demographic indicators, but also to service accessibility, reinforcing conditions of socio-territorial fragility [21].

Overall, the results pointed to the need for a progressive reconfiguration of service provision models toward forms of domiciliarity [35], where care and assistance are delivered closer to or directly within the living environments of the population. In contexts characterised by ageing, dispersion, and limited accessibility, domiciliary services emerge not as a complementary solution, but as a structurally necessary response to ensure continuity of care and territorial equity.

4.2 Energy System and Renewable Potential

The analysis of the energy dimension provided a complementary perspective [23] on the green transition, bringing to the forefront the issue of energy vulnerability as a key territorial condition. While the mapping of photovoltaic potential (Fig. 5), based on spatial analysis of exposure and built environment characteristics, revealed the presence of heterogeneous energy production capacities, this potential did not automatically translate into accessible or equitable energy provision [36].

The visualisation of energy data highlighted a significant misalignment between energy production potential, local demand, and service needs. Areas with higher renewable capacity are not necessarily those with higher population density or greater socio-assistential demand, resulting in asymmetries that contribute to conditions of energy vulnerability. This is particularly relevant in mountain and peripheral contexts, where infrastructural limitations, dispersed settlements, and limited accessibility can exacerbate both energy costs and service provision challenges.

In this framework, energy vulnerability [37] emerged not only as a function of income or consumption, but as a territorially embedded condition, shaped by the interaction between energy infrastructures, spatial accessibility, and socio-demographic dynamics. The integration of energy data with demographic and service-related information enabled the identification of contexts where limited access to energy resources overlaps with high social demand, reinforcing existing patterns of fragility.

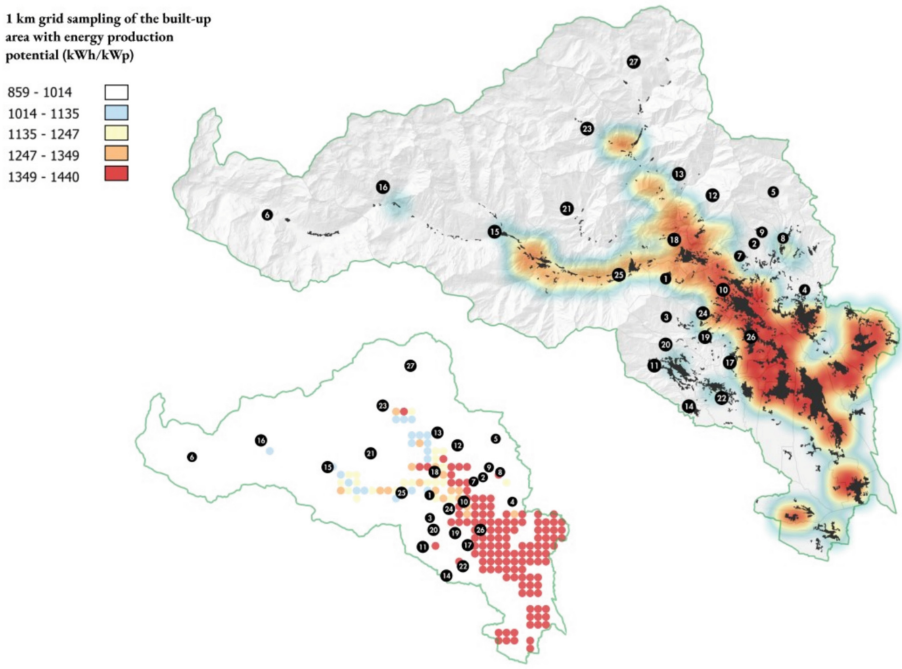


Fig. 5. Energy photovoltaic potential in the focus area (Authors' elaboration).

The presence of emerging initiatives such as the Fervores CERS suggested a potential pathway for addressing these challenges through decentralised and community-based energy systems. However, the analysis also revealed a significant gap between the conceptual potential of RECs and their current implementation. In particular, the development of these initiatives appears to be constrained by regulatory complexity, administrative delays, and the relatively recent activation of the supporting legislative framework. As a result, RECs have so far been only partially able to develop projects with tangible social impact, especially in relation to supporting vulnerable populations and socio-assistential services.

This limitation highlighted a broader issue in the current energy transition framework, where technical and regulatory advancements are not yet fully aligned with social and territorial needs. In the absence of targeted strategies, the risk is that renewable energy systems remain disconnected from local welfare systems, failing to address energy vulnerability in a structural way.

Overall, the results showed that the green transition, when analysed in isolation, tends to emphasize production potential while underestimating issues of accessibility, distribution, and social impact. The integration between green transition and grey transition is therefore essential to reframe energy systems as infrastructures with both environmental and social functions [24], capable of contributing to the reduction of territorial inequalities and the support of care systems [21], particularly in fragile and mountain areas.

4.3 SWOT Analysis Results

The SWOT analysis identifies territorial potentialities and critical vulnerabilities requiring strategic intervention. The findings concerning the energy dimension, demographic dynamics, and socio-assistential services are summarized in a matrix organized by three macro-themes:

- Territory dimension
- Socio-assistential services
- Energy dimension

To enable a more detailed analysis, each macro-theme was subdivided into specific sub-themes:

- Territory dimension: morphology, governance, population, mobility, and services.
- Socio-assistential services: aging and social needs, access, organizational structure and social relations.
- Energy dimension: general framework, production and consumption potential.

Each sub-theme revealed multiple strengths, weaknesses, opportunities, and threats. The data integrated into this matrix derived from a cross-analysis of energy, demographics, and social services, further enriched by three interviews with key stakeholders: Fervores CERS, the Green Community *Sinergie in Canavese* and CISS 38.

This integration of quantitative and qualitative data allowed a comprehensive comparison between desk analysis results and stakeholder insights.

The merged dataset reveals a complex and heterogeneous system characterized by strong interdependencies between territory, service provision, and energy potential. From a governance perspective, the presence of multi-actor cooperation (CERS Fervores, CISS 38, Green Community) represents a key enabling condition for integrated territorial development. However, this potential is constrained by administrative fragmentation and regulatory instability, which limit the scalability of coordinated interventions. However, the region suffered from morphological fragmentation, low population density, and an aging index far exceeding the regional average. Such conditions render service distribution uneven, particularly as travel times for public transit increased more sharply than for private vehicles as travel distance increases. While the CERS could act as a territorial aggregator and together with other key stakeholders could facilitate access to multi-level funding (EU, national, regional), administrative fragmentation and unstable planning instruments remained significant threats.

From a socio-assistential perspective, the expansion of the over-65 population increased healthcare demand, placing additional pressure on waiting lists and caregiver availability. The territory is characterized by a shortage of general practitioners; furthermore, 70% of the municipalities lack primary healthcare facilities. Initiatives like “Passi Montani” have addressed these gaps, while the upcoming “Prossimità” project aims to provide further reinforcement. There was a significant perceptual gap: social workers were often perceived as obstacles rather than support systems, unlike local administrators, who were individuals with whom the population directly expressed their care and assistance needs. Regarding energy, the primary weakness was the operational delay of Fervores CERS, alongside limited production potential in restricted zones. On the other

hand, strengths included the involvement of 10 new CER participants and the alignment between high-production areas and the largest town, such as Castellamonte. Currently, the most significant threat is the regulatory burden and bureaucracy imposed by the GSE.Overall, the SWOT synthesis suggests that the three domains cannot be interpreted independently, as they co-produce a unified framework of socio-territorial vulnerability. In this sense, territorial fragility, social demand pressure, and energy transition dynamics emerge as mutually reinforcing processes.

5 Conclusions

This paper has explored the relationship between green and grey transitions in rural areas through an integrated and place-based analytical framework, focusing on the North-Western Canavese area as an exploratory case study. The work represents a first step in investigating the interaction between demographic ageing, service contraction, and renewable energy potential within the ongoing SustAge project. While the results presented here are still preliminary, they provide the basis for further methodological developments and future operational applications that will be explored in the next phases of the project. The results confirmed the initial hypothesis that these transitions are not independent processes, but rather structurally interconnected dynamics that jointly shape territorial development.

The application of a multi-layered and interdisciplinary framework demonstrated the value of combining demographic, socio-assistential, infrastructural, and energy-related data within a shared analytical environment. This approach enabled a shift from sectoral interpretations to a systemic understanding of the territory, allowing identification of both synergies and mismatches between social demand, service accessibility, and energy potential. The use of a SWOT analysis provided a necessary bridge between hard data and field interviews, helping to contextualize territorial conditions within the larger framework of external opportunities and threats.From the perspective of the grey transition, the findings highlighted how demographic ageing is not only a demographic trend but a spatially embedded process, strongly influenced by territorial morphology, accessibility conditions, and service distribution. In the North-Western Canavese, some inequalities in access to socio-assistential services emerged, particularly in peripheral and mountain municipalities, where fragmented settlement patterns and infrastructural constraints limit the effectiveness of existing welfare systems. In this context, domiciliary and proximity-based care models emerged as a structurally necessary response to ensure continuity of care and territorial equity.

At the same time, the analysis of the green transition revealed that renewable energy potential, although present and in some cases significant, is unevenly distributed and not inherently aligned with local demand or socio-assistential needs. The results showed that energy systems, when considered in isolation, risk reinforcing existing territorial disparities, as areas with higher production potential do not necessarily correspond to areas of higher vulnerability or service demand. Energy vulnerability therefore emerged as a territorially embedded condition, shaped by the interaction between infrastructure, accessibility, and socio-demographic dynamics.

A key contribution of this study lies in the identification of spatial configurations in which overlaps between energy potential and social demand suggest opportunities

for integrated interventions. In particular, the presence of local initiatives such as RECs indicates a promising pathway for linking decentralised energy production with welfare systems and service provision. However, the current gap between the theoretical potential of these initiatives and their practical implementation highlights the need for more coherent regulatory, institutional, and governance frameworks.

Overall, the results suggested that addressing green and grey transitions requires moving beyond sectoral approaches toward integrated, place-based strategies that consider energy systems not only as technical infrastructures, but also as social infrastructures capable of supporting care, reducing inequalities, and enhancing territorial cohesion.

From a methodological perspective, the study demonstrated the central role of data visualisation and mapping as tools for integrating heterogeneous datasets and supporting interdisciplinary dialogue. Visual representations function as boundary objects, enabling different disciplinary perspectives to converge and facilitating the construction of shared knowledge.

Finally, this research represents a first step within the broader SustAge project. Future developments will focus on translating the identified synergies into operational strategies, exploring governance models, and testing pilot interventions aimed at integrating energy communities with socio-assistential services. Further research is also needed to refine the analytical framework, set a range of measurable indicators, and assess the replicability of the approach in other inner and rural contexts.

Acknowledgments. This research has been developed within the framework of the SustAge – Integrating green and grey transition project. The project is funded by the Italian Ministry of University and Research (MUR) under the Excellence Programme 2023–2027, through an internal call at the Law Department of the University of Turin also supported by Collegio Carlo Alberto. The project runs from January 2025 to December 2027 and is coordinated by Prof. Lorenza Mola.

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

Declaration of AI Usage. During the preparation of this manuscript, we used artificial intelligence-based tools to support language refinement, grammar and style improvement, and English editing. All outputs generated by these tools were critically reviewed, edited and approved by the author, who take full responsibility for the final content and language quality.

References

1. Bridge, G., Bouzarovski, S., Bradshaw, M., Eyre, N.: Geographies of energy transition: space, place and the low-carbon economy. Energy Policy **53**, 331–340 (2013). https://doi.org/10.1016/j.enpol.2012.10.066
2. Jenkins, K., McCauley, D., Heffron, R., Stephan, H., Rehner, R.: Energy justice: a conceptual review. Energy Res. Soc. Sci. **11**, 174–182 (2016). https://doi.org/10.1016/j.erss.2015.10.004

3. Commissione Europea: Il Green Deal europeo. (2019).
4. OECD: Rural Well-being: Geography of Opportunities. OECD Publishing (2020). https://doi.org/10.1787/d25cef80-en
5. Martinez-Fernandez, C., Audirac, I., Fol, S., Cunningham-Sabot, E.: Shrinking cities: urban challenges of globalization. Int. J. Urban Reg. Res. **36**, 213–225 (2012). https://doi.org/10.1111/J.1468-2427.2011.01092.X
6. McCann, P.: Perceptions of regional inequality and the geography of discontent: insights from the UK. Reg. Stud. **54**, 256–267 (2020). https://doi.org/10.1080/00343404.2019.1619928
7. Fikri, K.: Persistently poor, left-behind and chronically disconnected. Camb. J. Reg. Econ. Soc. **17**, 249–258 (2024). https://doi.org/10.1093/CJRES/RSAD048
8. Michael, J.A.B., Richter, M.: Demographic changes in Europe the rapid demographic shifts and the socioeconomic future of Europe Research Report. (2024)
9. Vitale Brovarone, E., Cotella, G.: Improving rural accessibility: a multilayer approach. Sustainability **12**, 2876 (2020). https://doi.org/10.3390/su12072876
10. Köhler, J., Geels, F.W., Kern, F., Markard, J., Onsongo, E., Wieczorek, A., Alkemade, F., Avelino, F., Bergek, A., Boons, F., Fünfschilling, L., Hess, D., Holtz, G., Hyysalo, S., Jenkins, K., Kivimaa, P., Martiskainen, M., McMeekin, A., Mühlemeier, M.S., Nykvist, B., Pel, B., Raven, R., Rohracher, H., Sandén, B., Schot, J., Sovacool, B., Turnheim, B., Welch, D., Wells, P.: An agenda for sustainability transitions research: state of the art and future directions. Environ. Innov. Soc. Transit. **31**, 1–32 (2019). https://doi.org/10.1016/J.EIST.2019.01.004
11. Basilico, P., Cartanese, G.M., D'Adamo, I., Ferrazza, M., Gastaldi, M., Uricchio, A.F.: Renewable energy communities and energy poverty: social and policy insights from empirical research in Italy. Energy Policy **210**, 115078 (2026). https://doi.org/10.1016/j.enpol.2026.115078
12. Boostani, P., Pellegrini-Masini, G., Klein, J.: The role of community energy schemes in reducing energy poverty and promoting social inclusion: a systematic literature review. Energies **17**, 3232 (2024). https://doi.org/10.3390/en17133232
13. SustAge - Integrating green and grey transition - Hub to Law. https://hubtolaw.it/projects/sustage-integrating-green-and-grey-transition/. Accessed 28 Mar 2026
14. Aulisio, A., Barbero, S., Barreca, A., Malavasi, G., Rolando, D.: From data collection to a cross-cutting analysis visualisation: territorial complexity overview to foster responsible tourism in rural areas. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., C. Rocha, A.M.A., Faginas Lago, M.N. (eds.) Computational Science and Its Applications – ICCSA 2024 Workshops. ICCSA 2024. Lecture Notes in Computer Science, vol. 14819, pp. 33–50. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65282-0_3
15. Pereno, A., Barbero, S.: Systemic design for territorial enhancement: an overview on design tools supporting socio-technical system innovation (2020). https://doi.org/10.4013/SDRJ.2020.132.02
16. Battistoni, C., Nohra, C.G., Barbero, S.: A systemic design method to approach future complex scenarios and research towards sustainability: a holistic diagnosis tool. Sustain. (Switz.) **11** (2019). https://doi.org/10.3390/su11164458
17. Lazzarini, G., Lucia, M.G.: Demographic aging: challenges and opportunities for ecological transition. J. Emerg. Perspect. **1**, 79–85 (2024). https://doi.org/10.36253/jep-16902
18. Martens, K.: Transport Justice. Transport Justice (2016). https://doi.org/10.4324/9781315746852
19. Moodie, J.R.: The key dimensions of just transitions - from concept to policy practice. J. Sustain. **1** (2025). https://doi.org/10.55845/jos-2025-1248
20. Skinner, M., Hanlon, N.: Ageing resource communities: new frontiers of rural population change, community development and voluntarism. Ageing Resour. Commun. New Front. Rural Popul. Change Commun. Dev. Voluntarism **56**, 1–219 (2015). https://doi.org/10.4324/9781315728209

21. Phillipson, C.: The 'elected' and the 'excluded': sociological perspectives on the experience of place and community in old age. Ageing Soc. **27**, 321–342 (2007). https://doi.org/10.1017/S0144686X06005629
22. Brueckner, J.K., Martin, R.W.: Spatial mismatch: an equilibrium analysis. Reg. Sci. Urban Econ. **27**, 693–714 (1997). https://doi.org/10.1016/s0166-0462(97)00004-5
23. Lund, P.D.: The link between political decision-making and energy options: assessing future role of renewable energy and energy efficiency in Finland. Energy **32**, 2271–2281 (2007). https://doi.org/10.1016/j.energy.2007.05.015
24. Baker, L., Newell, P., Phillips, J.: The political economy of energy transitions: the case of South Africa. New Polit. Econ. **19**, 791–818 (2014). https://doi.org/10.1080/13563467.2013.849674
25. Cerreta, M., Panaro, S.: Collaborative decision-making processes for local innovation: the CoULL methodology in living labs approach. GeoJournal Libr. **128**, 193–212 (2022). https://doi.org/10.1007/978-3-030-78536-9_12
26. Stankulova, A., Barreca, A., Rebaudengo, M., Rolando, D.: Emerging trends in the territorial and rural vulnerability-vibrancy evaluation. a bibliometric analysis. In: Gervasi, O., et al. Computational Science and Its Applications – ICCSA 2023 Workshops. ICCSA 2023. Lecture Notes in Computer Science, vol. 14106, pp. 277–288. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37111-0_20
27. Barreca, A., Curto, R., Rolando, D.: Assessing social and territorial vulnerability on real estate submarkets. Buildings **7** (2017). https://doi.org/10.3390/buildings7040094
28. Barreca, A., Curto, R., Malavasi, G., Rolando, D.: From data to knowledge by exploring weak socio-economic contexts and real estate market dynamics: the case of the Eporediese area (Italy). In: International Forum on Knowledge Asset Dynamics : Knowledge Drivers for Resilience and Transformation, pp. 1526–1540 (2023)
29. CISS38, https://www.ciss38.it/it-it/home. Accessed 24 June 2026
30. OECD, Union, E.: Social Economy in Europe: Contributing to Competitiveness and Prosperity. Local Economic and Employment Development (LEED) (2025). https://doi.org/10.1787/3432de93-en
31. Home | Green Community - Sinergie in Canavese. https://greencommunitysinergieincanavese.it/. Accessed 28 Mar 2026
32. COMUNITÀ ENERGETICA RINNOVABILE E SOLIDALE "CERS FERVORES." https://greencommunitysinergieincanavese.it/fervores. Accessed 28 Mar 2026
33. Sevaldson, B.: GIGA-mapping: visualisation for complexity and systems thinking in design. In: Nordes 2011: Making Design Matter (2011). https://doi.org/10.21606/nordes.2011.015
34. Lavecchia, L.: La povertà energetica in Italia. Prisma: economia, società, lavoro **2**(2024), 22–34 (2024). https://doi.org/10.3280/pri2024-002003
35. World report on ageing and health. https://www.who.int/publications/i/item/9789241565042. Accessed 28 Mar 2026
36. Calvert, K., Pearce, J.M., Mabee, W.E.: Toward renewable energy geo-information infrastructures: applications of GIScience and remote sensing that build institutional capacity. Renew. Sustain. Energy Rev. **18**, 416–429 (2013). https://doi.org/10.1016/j.rser.2012.10.024
37. Lavecchia, L., Miniaci, R., Valbonesi, P., Venkateswaran, G.: Energy poverty risk: a spatial index based on energy efficiency. Questioni di Economia e Finanza (Occasional Papers) (2024)

A Bi-Objective Model for Environmental Monitoring by Vessels and AUVs

Mahdi Moeini(✉)

ensIIE, SAMOVAR, 1 Place de la Résistance, 91000 Évry-Courcouronnes, France
mahdi.moeini@ensiie.fr

Abstract. In the context of environmental protection, this paper investigates offshore monitoring operations carried out by a primary vessel assisted by multiple autonomous submarines. Specifically, given a depot or port, a fleet consisting of a surface vessel equipped with several autonomous submarines is tasked with monitoring a predefined set of offshore locations. The fleet departs from the depot, completes the monitoring of all locations, and returns to the depot. Monitoring activities are performed through coordinated deployment of the surface vessel and the autonomous submarines, with the vessel acting as a mobile depot for launching and retrieving submarines via synchronized operations. The objective is to minimize total mission time, total energy consumption, or both simultaneously. To address this problem, we propose three mixed-integer linear programming (MILP) models and investigate the role of a bi-objective formulation. The models are implemented in Julia, tested on publicly available benchmark instances, and solved using standard optimization solvers. The results highlight the computational complexity levels as well as the trade-offs between minimizing mission time and reducing energy consumption.

Keywords: Vehicle Routing Problem · Maritime Logistics · Environmental Monitoring · Mixed-Integer Linear Programming · Heuristics

1 Introduction

Autonomous vehicles, e.g., drones, robots, etc., are used since several years for different objectives, e.g., parcel delivery, relief distribution, surveillance, environmental monitoring, etc. [4,11,19]. In this context, the use of autonomous boats, vessels, and submarines have particular importance due to the access restrictions to offshore zones [7,10]. Moreover, since oceans and seas cover about 71% of the Earth's surface, they require an effective protection, which can need frequent monitoring. The monitoring operations may be done for different military or civil purposes, e.g., discovering any potential leak of oil due to damage or sabotage actions. For this purpose, the boats, vessels, and submarines can be equipped with sensors or cameras to find the potential source of pollutants. Herein, an

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 503–519, 2027.
https://doi.org/10.1007/978-3-032-30494-0_32

efficient routing a fleet of maritime vehicles has a particular importance. Indeed, for example, through an optimized route planning, the detection of pollution can be done in a shorter time, which can help directly in environmental protection [20].

In the literature, we find several contributions that study routing a heterogeneous fleet of maritime vehicles for environmental monitoring. In particular, [20] investigates routing a fleet composed of a main vessel equipped with multiple identical submarines that can monitor (visit or cover) a given set of offshore locations in a minimal makespan. In this paper, we expand this concept to investigate alternative objectives, e.g., minimizing energy consumption of the fleet. The new models can be considered as alternative variants of the *multi-visit multi-drone covering salesperson problem with varying coverage* (mCSP-VC), where *varying coverage* refers approaching a location upto a certain distance to be close enough, without reaching the exact location, e.g., to witness, collect information, take pictures, and so on [20].

After presenting the *mixed-integer linear programming* (MILP) formulation of the problems, we solve them by the state-of-the-art solvers using publicly-available test instances and parameters. Moreover, we analyze the results of the models to investigate the benefits of using a bi-objective model taking into account both makespan and energy consumption.

The remainder of the paper is organized as follows. In Sect. 2, we have an overview on the existing literature related the problem studies in this paper. Section 3 is dedicated to the mathematical formulations of the problem. The solution methods are discussed in Sect. 4. The results of our computational experiments are reported in Sect. 5, and the last section is devoted to some concluding remarks.

2 Related Literature and Our Contributions

In terms of scientific research, many contributions have been developed in the context of autonomous vehicles, e.g., for parcel delivery, surveillance, environmental protection, etc. Detailed literature reviews can be found, e.g., in [4,8,20]. In this section, we present a concise review on some scientific articles that are closely related to the objectives of this paper.

In fact, in the literature, we observe different applications of the multi-visit multi-drone covering salesperson problem with varying coverage (mCSP-VC):

(i) Cooperative-based routing problems, where a fleet of autonomous vehicles (e.g., trucks combined with drones or robots) is involved with or without synchronization. In fact, there is a huge amount of research on routing drones [4,11,13], robots, or boats [10]. The general idea consists in routing a fleet of vehicles composed of one or more trucks, possibly equipped with a set of drones/robots. Furthermore, some studies consider a set of lockers or drone stations that are used as micro depots. In the literature, the typical objective consists in routing the vehicles to deliver parcels to all customers in minimal time. Some alternatives objective functions are also studied [19]. The

problems are formulated as MILP models that are then addressed either by a solver, or by an algorithm. In terms of solution methods, there are mostly heuristic or metaheuristic algorithms, but some exact algorithms have been developed too (see, e.g., [4] and references therein).

(ii) Coverage-based routing problems, where the objective consists in reaching a certain proximity of locations is sufficient have been generalized to involve multiple or autonomous vehicles (see, e.g., [2,3,20], and references therein). In particular, in [20], the fleet covering problem has varying coverage feature to visit or cover a given number of locations by a mixed fleet composed of a main vessel and multiple synchronized AUVs in minimal makespan.

(iii) offshore routing of maritime vehicles, where a combination of parts (i) and (ii) with some adjustments may be considered have particular importance, either in military or with civil objectives, e.g., tracking pollution, inspecting locations or under-water pipelines [15], or environmental monitoring [14].

In the context of maritime logistics, this paper investigates the use of a mixed fleet of vehicles composed of a main vessel equipped with a set of *autonomous underwater vehicles* (AUVs) (or submarines). The whole fleet starts its mission from a depot, visits or covers a given set of underwater spots, and returns to the depot. The AUVs are synchronized with the main surface vessel, such that each vehicle (main vessel or AUVs) can visit or cover one or multiple locations in a row. In the existing literature, the whole mission should be done in the minimal time [20]. We extend this existing concept through the following contributions:

- We consider alternative objective functions, i.e., minimizing energy consumption, which is aligned with environmental protection objectives. Due to the new objective function, the model becomes nonlinear, which is linearized by a classical technique.
- We introduce and investigate a bi-objective model which combines minimization of makespan and energy consumption.
- We conduct comparative computational study to investigate the role of alternative objective functions, e.g., minimizing energy consumption.

3 The Mathematical Models

In this section, we provide the mathematical formulations for the variants of the *multi-visit multi-drone covering salesperson problem with varying coverage* (mCSP-VC) presented in [20]. The detailed description of the mCSP-VC can be found in [20]; however, for the sake of completeness, we present it concisely in this section, and we use the corresponding notation for the two new models.

We consider a graph $G = (V, A)$, where $V = C \cup \{n + 1\}$ having n nodes $C := \{1, ..., n\}$ that should be visited or covered by either of the vehicles, the node $n + 1$ denotes the depot, and a set of arcs A that connect the nodes. The fleet of vehicles is composed of N_G vessels $L := \{1, ..., N_G\}$ such that each vessel is equipped with N_D identical autonomous submarines, also called *autonomous underwater vehicles* (AUVs).

We assume that the vessels have enough battery to visit all positions and return to the depot. However, each AUV has a limited endurance limit. Moreover, the vessel can launch and retrieve any AUV.

The set $K := \{1, ..., n-1\}$ indicates the index of all potential AUV sorties, where each sortie k in K consist of launch node of an AUV from a node, visiting or covering one or multiple nodes, and returns of the AUV back to the vessel.

When a vehicle, no matter the vessel or an AUV, remains on a node i, the range of covering by the vehicle is defined by the following formula [20]:

$$r_i = min(\theta, \alpha t_i + \beta), \tag{1}$$

where, θ is the maximal covering distance, $\alpha, \beta \geq 0$ are the estimated parameters, and t_i is the time spent by the vehicle on the node i.

A node j is considered as covered if

$$D_{ij} \leq r_i, \tag{2}$$

where, D_{ij} indicates the Euclidean distance between nodes i and j.

The further notation required for the formulation of the problems is summarized in Table 1.

3.1 The MILP Model Minimizing Makespan

In this section, we present the mathematical model of the mCSP-VC with the objective of minimizing the makespan, which is defined by the arrival time of the last vehicle to the depot. Mathematically, the objective function of the mCSP-VC is defined as follows:

$$\text{minimize } f(t) := \max\{t_{n+1}^{G,A}, t_{n+1}^{D,A}\} \tag{3}$$

The makespan defined by $f(t)$ is minimized subject to constraints (4)–(51), formulated as follows:

$$\sum_{k \in K} z_{ik}^{D} + z_i^{C} + z_i^{G} = 1, \forall i \in V \tag{4}$$

$$\sum_{(j,i) \in A} x_{jik} \geq z_{ik}^{D}, \forall i \in V, \forall k \in K \tag{5}$$

$$z_i^{G} = \sum_{(j,i) \in A} y_{ji}, \forall i \in V \tag{6}$$

$$h_{ik}^{R} + \sum_{(i,j) \in A} x_{ijk} = h_{ik}^{L} + \sum_{(j,i) \in A} x_{jik}, \forall i \in C, \forall k \in K \tag{7}$$

Table 1. Notations used for the mathematical models

Sets	
C	Set of all the nodes that must be covered or visited by the AUVs or the vessel.
V	Set of all the nodes, including the depot.
K	Index set of all possible AUV trips.
A	Set of arc among all nodes.
Parameters	
S_G	Speed of vessel.
S_D	Speed of AUVs.
D_{ij}	The Euclidean distance between nodes i and j.
$T_{ij}^G = D_{ij}/S_G$	The travel time of the vessel from node i to j.
$T_{ij}^D = D_{ij}/S_D$	The travel time of an AUV from node i to j.
B	The endurance limit of an AUV.
U	The energy consumption rate of an AUV per unit of time when moving.
$\tilde{U}$	The AUV energy consumption rate per unit of time when staying and taking measurements.
α	Radius expansion rate.
β	Intercept of radius expansion.
θ	Maximum coverage radius.
M	Sufficiently large positive number.
Variables	
x_{ijk}	Binary: Indicates whether the AUV trip k travels along arc $(i,j) \in A$.
y_{ij}	Binary: 1 if the vessel moves from node i to node j for $(i,j) \in A$, 0 otherwise.
z_i^G	Binary: Indicates whether node $i \in V$ is visited by the vessel.
z_{ik}^D	Binary: Indicates whether node $i \in V$ is visited by the AUV trip k.
z_i^C	Binary: Indicates whether node $i \in V$ is covered by staying in another node.
h_{ik}^L	Binary: Indicates whether node $i \in V$ is the launch node of AUV trip k.
h_{ik}^R	Binary: Indicates whether node $i \in V$ is the retrieval node of AUV trip k.
s_{ij}	Binary: 1 if by staying in node j, node i is covered where $(i,j) \in A$.
d_i	Integer: The number of onboard AUV on the vessel upon departure from node $i \in V$.
$t_i^{G,A}$	The vessel's arrival time at node $i \in V$.
$t_i^{G,L}$	The vessel's departure time from node $i \in V$.
$t_i^{D,A}$	The last AUV's arrival time at node $i \in V$.
$t_i^{D,L}$	The last AUV's departure time from node $i \in V$.
p_i^D	The remaining energy of the AUV at the point of departure from node i.
r_i	Radius of coverage at node $i \in C$

$$\sum_{(i,j)\in A} y_{ij} = \sum_{(j,i)\in A} y_{ji}, \forall i \in V \tag{8}$$

$$\sum_{i\in C} y_{i,n+1} = 1 \tag{9}$$

$$\sum_{j\in C} y_{n+1,j} = 1 \tag{10}$$

$$\sum_{(i,j)\in A} x_{ijk} \leq 1, \forall i \in V, \forall k \in K \tag{11}$$

$$\sum_{(j,i)\in A} x_{jik} \leq 1, \forall i \in V, \forall k \in K \tag{12}$$

$$\sum_{k\in K} x_{ijk} + \sum_{k\in K} x_{jik} \leq 1, \forall (i,j) \in A \tag{13}$$

$$z^D_{ik} + z^D_{jk} \geq x_{ijk}, \forall k \in K, \forall (i,j) \in A \tag{14}$$

$$h^L_{ik} \geq x_{ijk} + z^D_{jk} - z^D_{ik} - 1, \forall k \in K, \forall (i,j) \in A \tag{15}$$

$$h^R_{jk} \geq x_{ijk} + z^D_{ik} - z^D_{jk} - 1, \forall k \in K, \forall (i,j) \in A \tag{16}$$

$$\sum_{k\in K} h^L_{ik} \leq M z^G_i, \forall i \in V \tag{17}$$

$$\sum_{k\in K} h^R_{ik} \leq M z^G_i, \forall i \in V \tag{18}$$

$$h^R_{ik} + h^L_{ik} \leq 1, \forall i \in V, \forall k \in K \tag{19}$$

$$\sum_{i\in V} h^L_{ik} = \sum_{i\in V} h^R_{ik}, \forall k \in K \tag{20}$$

$$\sum_{i\in V} h^R_{ik} \leq 1, \forall k \in K \tag{21}$$

$$d_{n+1} + \sum_{j \in C} \sum_{k \in K} x_{n+1,j,k} = N_D \tag{22}$$

$$d_i + \sum_{k \in K} \sum_{(l,j) \in A} x_{ljk} \geq d_j + \sum_{k \in K} \sum_{(j,l) \in A} x_{jlk} - M(1 - y_{ij}), \forall (i,j) \in A \tag{23}$$

$$d_i + \sum_{k \in K} \sum_{(j,n+1) \in A} x_{j,n+1,k} \geq N_D - M(1 - y_{i,n+1}), \forall i \in C \tag{24}$$

$$p_i^D \geq B - M(1 - z_i^G), \forall i \in V \tag{25}$$

$$p_j^D \leq p_i^D - UT_{ij}^D - \tilde{U}(t_j^{D,L} - t_j^{D,A}) + M(3 - x_{ijk} - z_{ik}^D - z_{jk}^D), \forall k \in K, \forall (i,j) \in A \tag{26}$$

$$p_j^D \leq p_i^D - UT_{ij}^D - \tilde{U}(t_j^{D,L} - t_j^{D,A}) + M(3 - x_{ijk} - h_{ik}^L - z_{jk}^D), \forall k \in K, \forall (i,j) \in A \tag{27}$$

$$p_i^D \geq UT_{ij}^D - M(2 - x_{ijk} - h_{jk}^R), \forall k \in K, \forall (i,j) \in A \tag{28}$$

$$B \geq UT_{ij}^D - M(2 - x_{ijk} - h_{ik}^L), \forall k \in K, \forall (i,j) \in A \tag{29}$$

$$t_j^{G,A} \geq t_i^{G,L} + T_{ij}^G - M(1 - y_{ij}), \forall (i,j) \in A \tag{30}$$

$$t_i^{G,L} \geq t_i^{G,A} - M(1 - z_i^G), \forall i \in C \tag{31}$$

$$t_i^{G,L} \geq t_i^{D,A} - M(1 - h_{ik}^R), \forall k \in K, \forall i \in C \tag{32}$$

$$t_j^{D,A} \geq t_i^{D,L} + T_{ij}^D - M(1 - \sum_{k \in K} x_{ijk}), \forall (i,j) \in A \tag{33}$$

$$t_i^{D,L} \geq t_i^{D,A} - M(1 - \sum_{k \in K} z_{ik}^D), \forall i \in C \tag{34}$$

$$t_i^{D,L} \geq t_i^{G,L} - M(1 - h_{ik}^L), \forall k \in K, \forall i \in V \tag{35}$$

$$t_i^{D,L} \leq t_i^{G,L} + M(1 - h_{ik}^L), \forall k \in K, \forall i \in V \tag{36}$$

$$M \sum_{a \in C: a \leq i} z_{ak}^D \geq \sum_{a \in C: a \leq i} \sum_{b \in K: b \geq k} z_{ab}^D, \forall k \in K, \forall i \in C \tag{37}$$

$$r_j \leq \alpha(t_j^{D,L} - t_j^{D,A}) + \beta + M(1 - \sum_{k \in K} z_{jk}^D), \forall j \in C \tag{38}$$

$$r_j \leq \alpha(t_j^{G,L} - t_j^{G,A}) + \beta + M(1 - z_j^G), \forall j \in C \tag{39}$$

$$r_j \geq D_{ij} - M(1 - s_{ij}), \forall i \in V, \forall j \in C : (i,j) \in A \tag{40}$$

$$s_{ij} \leq \sum_{k \in K} z_{jk}^D + z_j^G, \forall i \in V, \forall j \in C : (i,j) \in A \tag{41}$$

$$\sum_{(i,j) \in A} s_{ij} \geq 1 - M(1 - z_i^C), \forall i \in C \tag{42}$$

$$\sum_{(i,j) \in A} s_{ij} \leq M z_i^C, \forall i \in C \tag{43}$$

$$x_{ijk} \in \{0,1\}, \forall (i,j) \in A, \forall k \in K \tag{44}$$

$$y_{ij}, s_{ij} \in \{0,1\}, \forall (i,j) \in A \tag{45}$$

$$z_i^G, z_i^C \in \{0,1\}, \forall i \in V \tag{46}$$

$$z_{ik}^D, h_{ik}^L, h_{ik}^R \in \{0,1\}, \forall i \in V, \forall k \in K \tag{47}$$

$$d_i \in \mathbb{N}^0, \forall i \in V \tag{48}$$

$$t_i^{G,A}, t_i^{G,L}, t_i^{D,A}, t_i^{D,L} \geq 0, \forall i \in V \tag{49}$$

$$0 \leq p_i^D \leq B, \forall i \in V \tag{50}$$

$$0 \leq r_i \leq \theta, \forall i \in C \tag{51}$$

The objective function (3) minimizes the makespan. Constraints (4)–(21) concern routing the fleet of the vehicles. Constraints (22)–(29) control the battery consumption and levels. Constraints (30)–(43) take care of the synchronization as well as the operational range of the vehicles. Finally, constraints (44)–(51) define the domain of the decision variables (for details, refer to [20]).

3.2 The MILP Model Minimizing the Consumed Energy

As an alternative objective, we are interested in minimizing the energy consumption during the trajectories, i.e., the total energy that vehicles consume while moving from one point to another. For this purpose, the objective function (3) is replaced by (52), in which the parameter κ is used as a coefficient to measure the rate of the energy consumption in the vessel versus the AUVs.

$$\begin{aligned} \text{minimize } g(x,y,z) := & \sum_{(i,j)\in A} \kappa \cdot U \cdot T^G_{ij} \cdot y_{ij} + \sum_{k\in K} \sum_{(i,j)\in A} U \cdot T^D_{ij} \cdot x_{ijk} \\ & + \sum_{k\in K} \sum_{j\in C} \cdot (t^{D,L}_j - t^{D,A}_j) \cdot z^D_{jk} \end{aligned} \tag{52}$$

Since the objective function (52) is nonlinear, we linearize it by introducing new variables as follows:

$$w^D_{jk} := (t^{D,L}_j - t^{D,A}_j) \cdot z^D_{jk} \quad : \quad \forall k \in K, \forall j \in C \tag{53}$$

Using the new variables, the objective function (52) becomes as follows:

$$\text{minimize } \sum_{(i,j)\in A} \kappa \cdot U \cdot T^G_{ij} \cdot y_{ij} + \sum_{k\in K} \sum_{(i,j)\in A} U \cdot T^D_{ij} \cdot x_{ijk} + \sum_{k\in K} \sum_{j\in C} w^D_{jk} \tag{54}$$

Moreover, to complete the formulation of the model, we add the following constraints to the model.

$$w^D_{jk} \leq M z^D_{jk}, \quad : \quad \forall k \in K, \forall j \in C \tag{55}$$

$$w^D_{jk} \geq 0, \quad : \quad \forall k \in K, \forall j \in C \tag{56}$$

$$w^D_{jk} \leq t^{D,L}_j - t^{D,A}_j, \quad : \quad \forall k \in K, \forall j \in C \tag{57}$$

$$w^D_{jk} \geq (t^{D,L}_j - t^{D,A}_j) - M(1 - z^D_{jk}), \quad : \quad \forall k \in K, \forall j \in C \tag{58}$$

The constraints of the model includes all constraints (4)–(51) as well as constraints (55), (56), (57), and (58).

By EC-mCSP-VC, we denote the mathematical model of the energy consumption minimizing mCSP-VC, which minimizes (54) subject to (4)–(51) and (55)–(58).

3.3 The Bi-Objective Model: Minimizing Makespan and Energy Consumption

From a bi-objective point of view, we can be interested in minimizing simultaneously makespan and energy consumption to investigate the relation and impact of each objective. This is motivated by searching for an optimal allocations and routing of heterogeneous fleet (vessel and AUVs) to decide which locations and in which order should be visited/covered by which vehicle type to minimize not only the mission time but also the energy consumption, which is meaningful when different vehicle types have different level of energy consumption.

For this purpose, we consider a third model, denoted by Bi-mCSP-VC, in which the objective function is composed of (3) and (54) that should be minimized simultaneously. The constraints of the model are the same as those of the second model, i.e., all constraints (4)–(51) as well as constraints (55), (56), (57), and (58).

Bi-mCSP-VC

$$\text{minimize } (f(t), g(x, y, z)) \tag{59}$$

$$\text{s.t. } (4) - (51) \text{ and } (55) - (58).$$

4 Methodology

We can use any standard MILP solver, e.g., Gurobi [6], to solve small instances of the single-objective MILP models presented in Sect. 3. Moreover, we present a simplified, yet efficient, version of the matheuristic presented in [20] to solve MILP models minimizing the makespan or energy consumption. In the scientific literature, heuristic, metaheuristic, and matheuristic algorithms are known as suitable choices for addressing complex optimization problems (see, e.g., [5,9,12,13,16–18,20], and references therein).

In the following, we present the description of the algorithm that we used to solve the single-objective models.

Algorithm 1 presents the pseudo-code of the matheuristic. The algorithm starts from a simple TSP tour in which each location is visited exactly once by the main vessel. Afterwards, a sequence of operators are applied to integrate the AUVs and to improve the quality of the solution. In the following, we describe different components of the algorithm.

Once the initial solution, i.e., a TSP tour, is available, the main loop of the algorithm begins and repeats a given number of iterations. Throughout this loop, first a random removal operator is applied by which either a given number of random locations are removed from the route, or a location is selected form a distance-based ordered list. Then, a randomly selected insertion operator is performed with the objective of inserting a vessel or AUV to accomplish a visit or cover a location. Two insertion operators are selected randomly: *covered insertion*

and *greedy insertion*. The former one checks if the removed location could be covered by a vehicle. The latter operator verifies all possibilities to see if an insertion in the vessel or AUV route gives a better solution to perform the operation.

The obtained solution is passed to the *one-step* sub-procedure (line 7 of Algorithm 1). This step of the algorithm, along with the new solution, takes a given number of parameters, in particular an upper bound of the objective value of the current solution, then builds a restricted MILP model in which some parts of the model are fixed. More precisely, the solution generated by the vessel's and AUVs' routes are the fixed parts of the model to which an upper bound is set. Afterwards, the restricted MILP model is solved by the MILP solver under a shortened time limit. If the restricted model is feasible, its solution is then evaluated versus the current best solution to replace it if the new one has a better objective value.

Algorithm 1. The matheuristic for the MILP models presented in Section 3.

1: Read the instance and initialize the parameters
2: Generate a TSP tour s
3: **for** iterations $\leq$ maximum number of iterations **do**
4: Select randomly a removal operator and apply it
5: Select randomly a insertion operator and apply it
6: Compute an upper bound obj_{ub}
7: Apply the *one-step* procedure to generate s_{new}
8: **if** s_{new} is better than the current best solution s_c **then**
9: $s_c \leftarrow s_{new}$
10: **end if**
11: **end for**
12: **return** s_c

Figure 1 gives sample visualizations of the solutions provided by the models minimizing makespan and energy consumption (with $\kappa = 2$). As we observe, when it is about minimizing the total mission time, parallel dispatching of vehicles is used, while minimizing the energy consumption avoid such a case.

5 Computational Experiments

This section is dedicated to the description of the computational experiments and their numerical results.

5.1 Settings of the Experiments

The MILP models presented in Sect. 3 are solved by the existing solvers and tested on publicly available instances with up to 10 nodes. More precisely, the

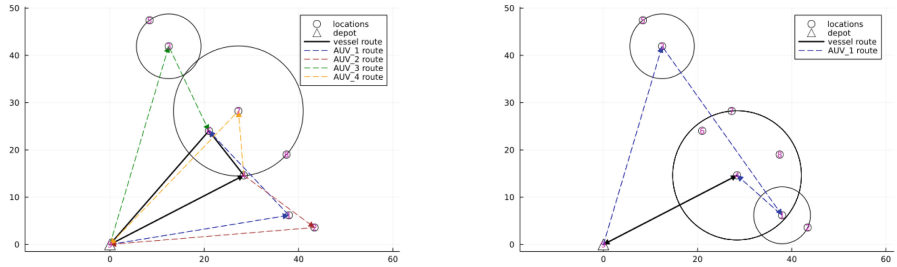


Fig. 1. Sample results of two models on a same instance. Left: the sample output of the model minimizing the makespan. Right: the sample output of the model minimizing the energy consumption with $\kappa = 2$.

MILP models are implemented in Julia [1]. The single-objective models are solved by the standard solver Gurobi (with time limit of 1800 seconds), while the bi-objective model is solved by Jump package of Julia. For the parameters of the models, we use the same values that we find in the literature [20]. Specifically, $B = 400.0$ (the endurance limit of an AUV), $U = 2.0$ (the energy consumption rate of an AUV per unit of time when moving), $\tilde{U} = 1.0$ (AUV energy consumption rate per unit of time when staying and taking measurements), $\alpha = 1.5$ (the radius expansion rate), $\beta = 0.0$ (the intercept of radius expansion), $\theta = 15.0$ (the maximum coverage radius), $S_D = 1.25$ (speed of AUVs), $S_G = 1.0$ (speed of vessel), $ND = 4$ (the number of AUVs).

Finally, for the model minimizing the energy consumption as well as the bi-objective model, we set the value of $\kappa = 2$, which indicates that the vessel consumes twice as an AUV per unit traveled distance.

5.2 Numerical Results

The following figures visualize the results of solving the three presented models for different instances and the corresponding parameters as described in Sect. 5.1.

On Figs. 2, the left-side figure shows obtained by the solver, i.e., the values of the consumed energy by the vehicles measured by the models. The focus of the first model (mCSP-VC) is on minimizing the makespan while the other models take into account the energetic (and consequently, ecologic) impact of the vehicles too. Consequently, the mCSP-VC model gives results that are not so ecological as the results provided by the two other models. Moreover, the values obtained by the bi-objective model coincide with the values obtained though the model minimizing the energy consumption, i.e., EC-mCSP-VC. In these experiments, when required, the value of κ is set to 2. The right-side figure shows similar experiments, but the single-objective models are solved by the metaheuristic presented in Sect. 4. As we observe in Figs. 2, the metaheuristic algorithm provides high-quality solutions for both the mCSP-VC as well as the EC-mCSP-VC models.

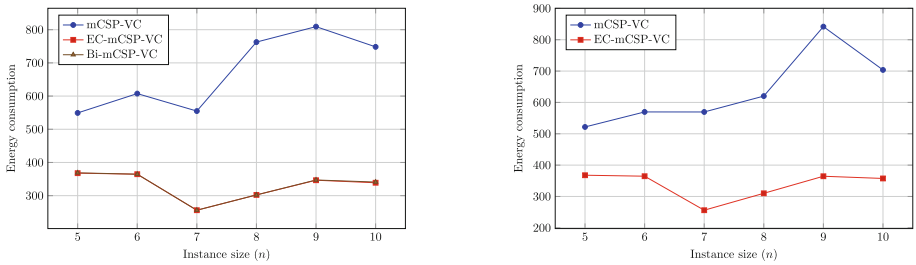


Fig. 2. Energy consumption values measured by three models obtained by using the solvers (Left) versus the results of the matheuristic (Right). In case of using the solvers, the values obtained by the bi-objective model coincide with the values obtained though the model minimizing the energy consumption, where we set $\kappa = 2$. The matheuristic is designed and used for solving only the single-objective models.

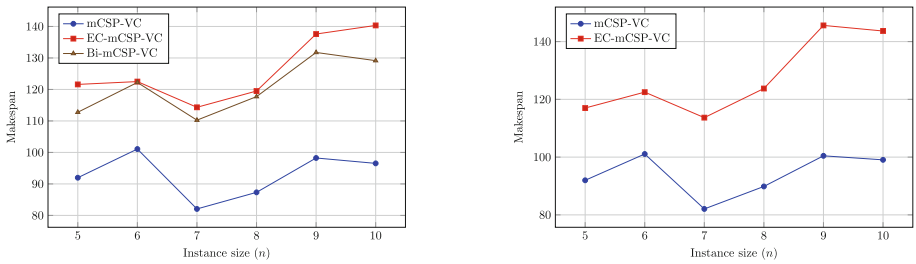


Fig. 3. Makespan values measured by three models obtained by using the solvers (Left) versus the results of the matheuristic (Right). In the corresponding models, we set $\kappa = 2$. The matheuristic is designed and used for solving only the single-objective models.

On Figs. 3, the left-side figure presents the results of the solvers in addressing the three models. In these results, the focus on the values of the makespan. The focus of the first model is on minimizing the makespan which is not fully the case of the other models, in particular the bi-objective model Bi-mCSP-VC considers a combination of makespan and energy consumption. Thus, as we observe, there is a big gap between the values obtained by the model minimizing the makespan versus the others. In these experiments, whenever required, the value of κ is set to 2. On the right-side figure, we present the results of the similar experiments where the single-objective models are addressed by the metaheuristic presented in Sect. 4. Similar to the case of Figs. 2, we observe also in Figs. 3 that the metaheuristic algorithm provides high-quality solutions.

Figure 4 shows the computation time required by the solvers and the matheuristic in addressing the three models. As we observe, solving the energy consumption model (EC-mCSP-VC) is fast, and using a heuristic approach doesn't seem to be necessary. Indeed, most of the tested instances are solved in less than one minute by the solver Gurobi. In this regard, the matheuristic

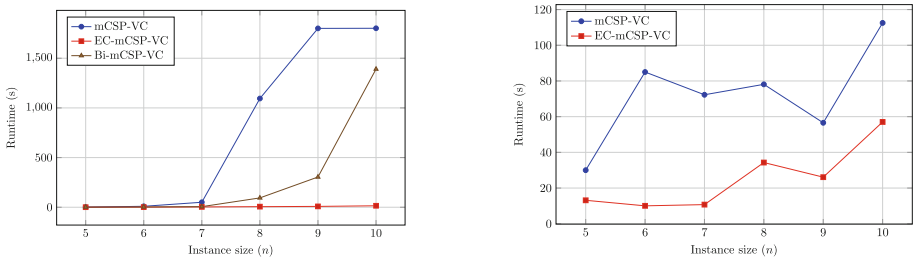


Fig. 4. Required computation time for solving the single- and bi-objective models. The figure on the left side corresponds to the results of the solvers, and the right-side figure illustrates the computation time of the matheuristics for solving the single-objective models.

finds high-quality solutions, but in many cases, it requires longer computation time than the solver.

However, when dealing with the bi-objective model, solving the model remains time consuming, even though it is still less challenging that the model minimizing the makespan, i.e., the mCSP-VC.

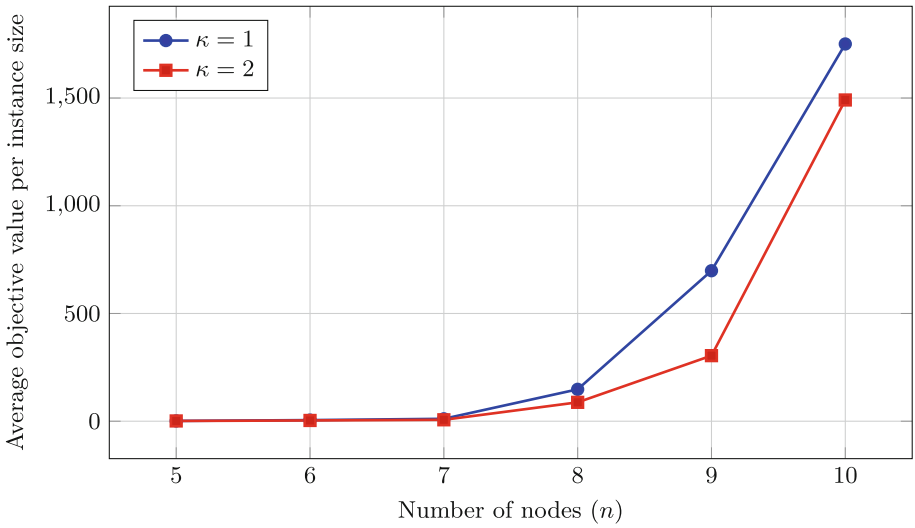


Fig. 5. The bi-objective model Bi-mCSP-VC: Objective value as a function of the problem size n for two values of κ.

Figure 5 illustrates the computation time required to solve the bi-objective model for $\kappa \in \{1, 2\}$. As we observe the computation time increases fast and the growing rate of computation time is reduced by increasing the value of κ.

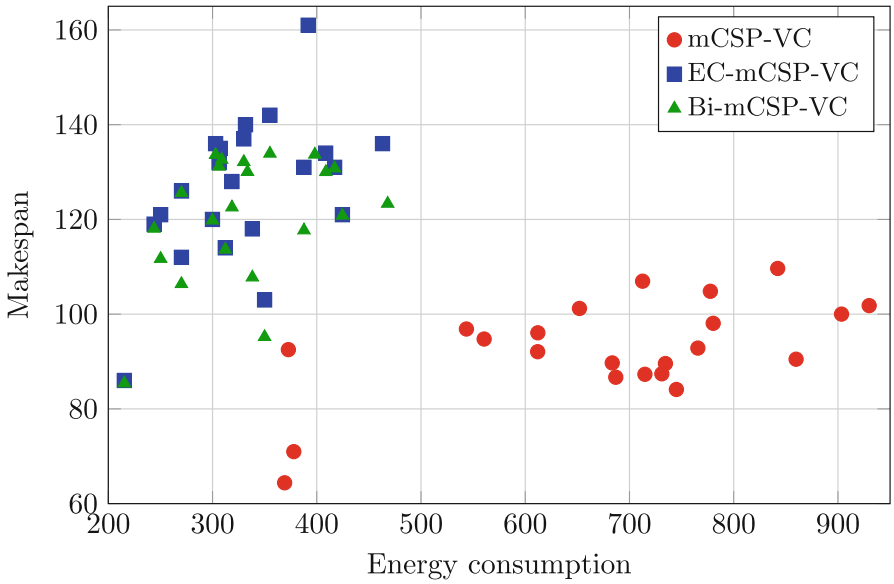


Fig. 6. Benchmarking makespan versus energy consumption levels provided by the solutions of the single- and b-objective models.

Finally, Fig. 6 provides an overview on the results of the models. More precisely, the figure shows how the objective values of the three models are scattered in a comparison of makespan values versus energy consumption levels. As expected, the single-objective models give quite different measures of makespan and energy consumption. Meanwhile, the bi-objective model provides a compromise between these objectives, and sometimes its values coincide with those of the single-objective model minimizing the energy consumption.

6 Conclusion

In this paper, we introduced two new MILP models that deal with minimizing energy consumption of a heterogeneous fleet composed of a vessel and AUVs with the mission of visiting or covering a set of locations. These are new variants of an existing model [20], named mCSP-VC. In particular, we present a bi-objective model that minimizes simultaneously makespan and energy consumption. We solved all models either by a solver or a matheuristic. The results of computational experiments on publicly available instances are reported. Moreover, we observe that the models provide solutions that can be quite different from the solution of the mCSP-VC.

The future research directions consist in developing efficient algorithms for the bi-objective model as well as extending the models for multiple vessels. The research in these directions is in progress, and the results will be reported in the future.

Acknowledgements. The author would like to thank Mr. Grégoire BROUARD for his assistance and help as well as the anonymous reviewers for their generous and constructive feedback.

References

1. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. SIAM Rev. **59**(1), 65–98 (2017)
2. Biswas, A., Babaee Tirkolaee, E., De Narayan, L., Yu, V.F., Pal, T.: On multi-objective multi-coverage covering salesman problem. Soft. Comput. **29**, 6005–6022 (2025)
3. Di Placido, A., Archetti, C., Cerrone, C., Golden, B.: The generalized close enough traveling salesman problem. Eur. J. Oper. Res. **310**(3), 974–991 (2023)
4. Enes Cengiz, C.Y., Kahraman, H.T.: A literature review of drone-truck routing problems: challenges and future research directions. RAIRO Oper. Res. **59**(5), 3169–3205 (2025)
5. Goeke, D., Moeini, M., Poganiuch, D.: A variable neighborhood search heuristic for the maximum ratio clique problem. Comput. Oper. Res. **87**, 283–291 (2017)
6. Gurobi: Gurobi Optimizer Reference Manual (2019)
7. Liu, Y., Xia, J., Xu, Z.: Drone scheduling optimization for continuous sea area monitoring. Transp. Res. Part E Logist. Transp. Rev. **208** (2026). https://doi.org/10.1016/j.tre.2026.104701
8. Macrina, G., Pugliese, L.D.P., Guerriero, F., Laporte, G.: Drone-aided routing: a literature review. Transp. Res. Part C Emerg. Technol. **120**, 102762 (2020)
9. Moeini, M.: Optimizing airline pilots training plans: a mixed integer linear programming approach. Commun. Comput. Inf. Sci. **2144**, 289–301 (2024)
10. Moeini, M.: Routing a truck and its autonomous supporting vehicles in humanitarian relief distribution. In: 26th congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF), pp. 1–2. ROADEF (2025)
11. Moeini, M., Le, D.T.D.: Multiple traveling salesman problem with a drone station: using multi-package payload compartments. Commun. Comput. Inf. Sci. **2144**, 226–237 (2024)
12. Moeini, M., Schermer, D., Wendt, O.: Cascading-tree algorithm for the 0-1 knapsack problem (in memory of Heiner Müller-Merbach, a former president of IFORS), pp. 1–11 (2024)
13. Moeini, M., Wendt, O., Schummer, M.: A bi-objective routing problem with trucks and drones: minimizing mission time and energy consumption. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2023 Workshops, ICCSA 2023. LNCS, vol. 14106, pp. 291–308. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37111-0_21
14. Niu, H., Adams, S., Lee, K., Husain, T., Bose, N.: Applications of autonomous underwater vehicles in offshore petroleum industry environmental effects monitoring. J. Can. Pet. Technol. **48**(5), 12–16 (2009)
15. Sahoo, A., Dwivedy, S.K., Robi, P.: Advancements in the field of autonomous underwater vehicle. Ocean Eng. **181**, 145–160 (2019)
16. Schermer, D., Moeini, M., Wendt, O.: Algorithms for solving the vehicle routing problem with drones. In: Nguyen, N.T., Hoang, D.H., Hong, T.-P., Pham, H., Trawiński, B. (eds.) ACIIDS 2018. LNCS (LNAI), vol. 10751, pp. 352–361. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75417-8_33

17. Schermer, D., Moeini, M., Wendt, O.: A hybrid VNS/Tabu search algorithm for solving the vehicle routing problem with drones and en route operations. Comput. Oper. Res. **109**, 134–158 (2019)
18. Schermer, D., Moeini, M., Wendt, O.: A matheuristic for the vehicle routing problem with drones and its variants. Transp. Res. Part C Emerg. Technol. **106**, 166–204 (2019)
19. Schermer, D., Moeini, M., Wendt, O.: The drone-assisted traveling salesman problem with robot stations. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 1308–1317 (2020)
20. Torabi, P., Hemmati, A.: A synchronized vessel and autonomous vehicle model for environmental monitoring: mixed integer linear programming model and adaptive matheuristic. Comput. Oper. Res. **183**, 108389 (2025)

The Effects of the Digital Transition on the Environment and Human Well-Being. From Cognitive and Evaluative Protocols to Possible Actions for Regional Planning

Fabrizio Battisti(✉) and Melania Perdonò

Department of Architecture, University of Florence, 50121 Florence, Italy
{fabrizio.battisti,melania.perdono}@unifi.it

Abstract. The digital transition represents one of the most impactful processes of transformation in contemporary society, capable of producing dual effects (positive and negative). This contribution, which is part of the PRIN 2022 PNRR project titled "Housing mobility and digital transition: assessment tools and technologies to understand current and future housing needs of people, supporting governance processes and territorial regeneration, starting from the analysis of the impacts of the digital transition," first proposes a systematisation of the normative (hard) and cultural (soft) methodological approaches for managing the digital transition. It then offers an integrated governance proposal that combines normative (hard) and cultural (soft) tools to support regional planning and the consequent housing policies potentially induced by the digital transition.

Keywords: Digital Transition · environmental sustainability · human well-being · Territorial governance · Urban planning

1 Introduction

Digital transformation represents one of the most significant processes of change in the 21st century. It includes the widespread adoption of digital technologies such as artificial intelligence, the Internet of Things (IoT), cloud computing, big data, blockchain, 5G/6G networks, digital platforms, and advanced automation systems. This shift, accelerated over the past two decades, has impacted manufacturing sectors, public services, education, healthcare, finance, and social relationships, leading to profound changes in everyday human life [1, 2].

Meanwhile, the global society faces two systemic challenges: the environmental crisis and the deterioration of many aspects of human well-being, including mental health, socio-economic inequalities, and job precariousness [3, 4].

The shift of traditionally physical functions to digital platforms – work, commerce, public services, communication, entertainment, education – is often described as a process of 'dematerialisation'. However, recent literature encourages distinguishing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 520–537, 2027.
https://doi.org/10.1007/978-3-032-30494-0_33

between apparent dematerialisation and invisible material reconfiguration [5]. Digitalisation sometimes does not eliminate matter; rather, it reorganises it in space and within value chains. This has effects on both the environment and society. Its environmental impacts need to be analysed on multiple levels: CO_2 emissions, traffic, land use, impacts on the atmosphere, hydrosphere, and subsoil. It is necessary to understand these effects. Several studies recognise that digitalisation can replace some physical activities, with potential positive effects on the environment [6–8]. 'Dematerialisation' – that is, the replacement of physical goods and services with digital equivalents—has, for example, reduced the need for physical transportation of people and materials, with potential positive impacts on CO_2 emissions associated with urban and interurban traffic [9].

Furthermore, digitalisation has profoundly transformed daily routines: work, relationships, consumption, learning, leisure. The ability to carry out activities online reduces spatial constraints but alters the perception of time [10]. Physical space can lose its functional centrality and its social value. These are changes that have deep environmental impacts: smart working reduces traffic but modifies domestic energy consumption [11]; e-commerce changes urban logistics and consumption habits [12]; administrative digitalisation reduces paper use but increases ICT infrastructure; quality of life depends on the ability to integrate environmental efficiency, psychological well-being, and social equity [13].

In light of the issues outlined above, it is necessary to establish how to manage changes in the physical spaces of cities and to understand the related changes in residential patterns. The research "Housing Mobility and Digital Transition: Tools and Evaluation Technologies to Understand Current and Future Housing Needs of People, Supporting Governance and Territorial Regeneration Processes," funded under the PRIN 2022 PNRR, examined how digital transition, by effectively altering customs and ways of life, can open new possibilities in terms of choosing where to live. This contribution, within the scope of the research and subject to experimentation in the metropolitan area of Florence, explores the intersections among digitalisation, environmental sustainability, and quality of life. It highlights the effects of digital transition on the environment (both positive and negative), with the aim of outlining methodological approaches and actions—both regulatory (hard) and cultural (soft)—to turn the impact of digital transition on a specific theme, urban policies, and territorial governance into an opportunity rather than a threat.

The operational result of the contribution is a proposal/protocol for assessing the impact of the digital transition on housing mobility, aimed at supporting public decision-making in regional planning processes. The proposed approach aims to provide an interpretative and operational framework for urban planning and territorial governance, as well as housing policies, that supports a more sustainable, inclusive, and equitable digital transition.

Below is a proposed literary review (paragraph 2) aimed at: i) identifying the effects of the digital transition in terms of systemic interaction between the environmental dimension and well-being within the framework of digital transformation; ii) exploring the opportunities offered by the digital transition in terms of reorganising settlement functions from a digital plane to a material plane. Subsequently, a procedure (paragraph 3) is proposed to validate the experimentation of the aforementioned PRIN 2022

PNRR Research, with the aim of defining the methodological approaches preparatory to the identification of key elements for 'managing/governing' the digital transition in the settlement context, with reference to residential functions. Following this, an operational proposal (paragraph 4) is developed, integrating possible hard and soft actions for regional planning activities. Finally, the conclusions (paragraph 5) of this work are drawn.

2 Literary Review: The Effects of the Digital Transition

Digital transition, understood as the spread of information technology, communication, and the automation of production processes in services and daily practices, is frequently associated with a process of modernisation deemed positive, capable of bringing significant benefits to the global community as a whole. However, various studies and research over the past fifteen years have analysed, alongside the positive aspects of the topic, also the negative ones, highlighting how digital transition is not constituted in an immaterial realm, but rather consists of an important material technological-social network: just think of data centres, telecommunication networks, logistics chains, etc., which must necessarily be created for the proper utilisation of digital technology [14–16]. This, of course, has positive impacts by reducing travel and time, but can sometimes have environmental repercussions. Furthermore, digitalisation particularly affects human well-being and individuals' everyday experiences: in this case as well, while it brings enormous advantages (such as easy access to services and information, reduced travel, and better organisational work), it can also negatively influence aspects related to social relationships, which tend towards isolation, and consequently impact mental and physical health [17–20].

The following review summarises recent literature on the impacts of the digital transition, analysing both the environmental effects resulting from the digitalisation of physical functions and those on human well-being related to people's daily experiences. The main objective is to define the state of the art regarding the positive and negative effects of the digital transition on the environment and quality of life.

2.1 Environmental Impacts of the Digital Transition

A significant body of literature interprets digitalisation as an enabling infrastructure for decarbonisation and the efficient use of resources, thereby supporting the ecological transition [21]. More generally, recent literature tends to interpret this phenomenon within the framework of the so-called "Twin Transition", emphasising how digital and ecological transitions can mutually reinforce each other only in the presence of appropriate governance and regulatory policies [22]. Numerous factors define positive environmental effects; for example, devices capable of monitoring and optimising energy consumption in real time (IoT, sensors, advanced analytics, etc.) allow the identification of operational inefficiencies, thereby reducing waste and improving the allocation of energy resources, especially in industrial contexts responsible for the highest energy consumption globally [23, 24]. This dynamic is often linked to the Industry 4.0 paradigm, which describes the interaction between cyber-physical systems, IoT sensors, data analysis, and advanced

automation in manufacturing processes, aiming to improve operational efficiency and resource management [25]. Furthermore, several studies emphasise that assessing the actual environmental benefits of these solutions requires a life-cycle assessment (LCA) approach that considers not only the use phase but also the production, maintenance, and end-of-life of digital technologies. Equally relevant is the theme of dematerialisation, understood as the substitution of goods, supports, and material activities with digital elements, a potentially environmentally sustainable factor. This phenomenon leads to the so-called "substitution effect": for example, the introduction of digital databases and archives increasingly impacts positively on reducing paper consumption, bringing significant environmental benefits (such as less deforestation); additionally, the spread of online public services (digitalisation of public administration), telemedicine, smart working, etc., is gradually decreasing CO2 emissions, with positive impacts on environmental pressure [26]. On the other hand, the dematerialisation effect can also be interpreted differently: Coroama, Moberg, and Hilty (2015) [27] demonstrate that electronic media can be a more energy-efficient alternative to some traditional activities, such as long-distance travel or certain physical supports. However, the same authors emphasise that these technologies can integrate with existing practices rather than replace them. In such cases, the effect is not a reduction in impact but an overall increase (rebound effect). This dynamic is often understood considering the so-called Jevons Paradox, which states that efficiency improvements can lead to an overall increase in consumption rather than a decrease. For this reason, the availability of compact and efficient devices does not guarantee dematerialisation: the entire lifecycle of media systems must be considered, especially the macro-level effects on the total volume of production and consumption. Another consideration concerns the themes of circular economy and traceability: digital technologies, such as supply chain information systems and, in some interpretations, blockchain, are proposed as tools to increase transparency along the supply chain and improve the monitoring of product lifecycle [28, 29]. Traceability can make reuse strategies more effective and efficient, increasing producer responsibility and thus facilitating environmental impact analysis and compliance with standards. The urban planning sector is also not immune to the new technologies arising from the digital transition; in particular, smart city models contribute positively to reducing emissions, urban transport networks, water supply, and waste disposal, resulting in a favourable environmental impact. These processes are further accelerated by the introduction of Artificial Intelligence (AI), which requires high computational capacity. While adopting these models positively impacts the environment, it also entails significant demand for digital infrastructure, hardware, and devices, which, in turn, increases energy consumption for their operation and contributes to increased electronic waste (e-waste). Digital materials are heavily dependent on critical minerals and rare-earth elements (such as lithium, cobalt, and neodymium), which are necessary for producing batteries, smartphones, servers, and other electronic infrastructure. These activities are sometimes responsible for current deforestation, which negatively affects biodiversity and contaminates soil and groundwater. Added to this is the well-known issue of electronic waste disposal, a rapidly growing global flow that, if not properly managed, risks causing serious environmental and health problems. Moreover, the energy-intensive use of AI systems, driven by the high energy consumption of data centres, is increasing sharply, with a peak expected

by 2030 [30]. Additionally, the ICT sector is currently estimated to be responsible for between 2% and 4% of global greenhouse gas emissions, an increase attributable to the expansion of data centres, transmission networks, and computational loads associated with Big Data and Artificial Intelligence [31, 32]. Furthermore, a factor not to overlook is that these systems require a large volume of water for server cooling (see evaporative cooling systems), which can increase water stress. Although studies on the energy efficiency of data centres have improved over time, progress in this area has sometimes been offset by increased data traffic and computational demand [33].

The production of sensors, IoT devices, and data centres involves a significant carbon footprint, sometimes increasing energy demand rather than reducing it. Additionally, many of the environmental costs of digitalisation are unevenly distributed along global value chains: while a substantial part of the benefits is concentrated in highly digitalised countries, numerous material, extractive, and disposal impacts fall on the countries of the Global South, raising issues of environmental and digital justice [34]. It is therefore clear that if such policies are not supported by appropriate green energy infrastructure and circular economy processes, smart cities risk producing unintended effects [30, 35, 36].

To summarise the main environmental effects of the digital transition identified in the analysed literature, Table 1 presents the main areas of application and their associated positive and negative effects.

Table 1. The effects of the digital transition on the environment

TOPIC	POSITIVE EFFECTS	NEGATIVE EFFECTS
Energy efficiency	• Monitoring of consumption; • Reduction of waste; • Optimisation of production processes	• Increase in the necessary digital infrastructure; • Increase in computational consumption
Dematerialization	• Reduction of paper use; • Minor physical supports; • Fewer movements	• Rebound effect; • Increase in digital consumption
Circular Economy and traceability	• Product lifecycle monitoring; • Supply chain transparency	• Technological complexity; • Increase in the use of digital infrastructure
Smart cities	• Transport optimisation; • Efficient management of services	• Increase in energy demand; • Increase in connected devices
AI and digital infrastructure	• Better management of urban and industrial systems	• High energy consumption (data centre, e-waste)

Overall, the literature shows that the digital transition presents ambivalent environmental effects: while it can contribute to greater efficiency in resource use and the reduction of certain environmental pressures, it also entails new impacts related to the energy consumption of digital infrastructure, resource extraction, and electronic waste

production, underscoring the need for sustainable governance. From this perspective, the environmental impact of digitalisation cannot be assessed solely in terms of local efficiency; instead, it must be analysed considering the entire socio-technical system in which the technologies are integrated, including the modes of production, use, and final disposal.

2.2 Impacts of the Digital Transition on Human Well-Being

As anticipated in the introductory section, the digital transition not only impacts the environmental level but also profoundly affects the social dimension and that of individuals, sometimes significantly altering daily practices (socialisation, work, education, etc.). Numerous studies on smart working and remote work highlight contrasting impacts: on one hand, this new productive modality offers greater flexibility and autonomy in work practices, balancing work and private life with potential positive effects on job satisfaction [37]; on the other hand, phenomena of blurring between the two spaces (work and private) can contribute to burnout and stress, preventing clear temporal boundaries between work activities and rest. This concept can be summarised in the paradigm of "work-life flow", recently introduced by Joanna Wells and Fiona Scheibein (2023) [38]. Furthermore, particularly in remote work, there is a high likelihood of developing professional isolation (loneliness paradox) [19], understood as disconnection from the social network of work, composed of colleagues, clients, etc., with possible negative effects on the sense of belonging and psychological well-being. The lack of physical contact and nonverbal signals can reduce oxytocin production, the social bonding hormone. However, there is no evidence that this influences emotional attachment to the organisation [39]. More generally, some studies show that excessive digital connections, whether work-related or leisure-related, can be associated with increased stress, reduced concentration, and mental fatigue [40, 41]. Intensive use of digital technologies can indeed generate phenomena of information overload and technostress, linked to the constant availability of information, notifications, and attention requests. Another branch of the literature concerns the impact of social platforms on psychological well-being. Several studies highlight that intensive use of social media can be associated with increased perceived loneliness, anxiety, and depressive symptoms, especially among adolescents and young adults [42, 43]. In this context, the phenomenon of Fear of Missing Out (FOMO) and continuous social comparison mediated by algorithms can reduce subjective life satisfaction and promote compulsive connection practices [44, 45].

Similarly, the increasingly immediate access to online shopping platforms operates on a dual track: on one hand, it contributes to more efficient time management and can reduce physical travel related to purchasing goods; on the other hand, it can generate potential forms of compulsive consumption and accelerate technological replacement cycles [46], contributing to the mass production of objects and services, with significant implications not only in terms of energy consumption and carbon footprint (see paragraph 2.1), but also in persistent dissatisfaction and possible conditions of stress, psychological discomfort, and depressive symptoms. This new culture of online shopping can also induce behaviours such as sedentariness and micro-movements, and some studies have associated this new way of living daily life with a primary driver of type 2 diabetes.

Online connectivity, however, especially in contexts where digital access represents an indicator of social inclusion, can be correlated with subjective well-being [47]. It is evident that, conversely, the so-called digital divide phenomenon is developing, highlighting how minority social groups (the elderly, populations with fewer technological skills, etc.) can sometimes be completely excluded from various dynamics and the benefits derived from the use of online tools, thereby increasingly exacerbating socio-economic inequalities [48]. Digitalisation has also generated positive effects on educational and informational offerings, contributing to the dissemination of new learning opportunities and the gradual democratisation of access to training through the availability of digital educational resources, online courses, and open learning platforms [49]. Not only do some studies show that students who have taken online courses achieve slightly better results than those attending in-person courses [50]. Finally, digitalisation has contributed to the spread of new forms of work organisation based on digital platforms, often referred to as the gig economy. Although these models offer new income opportunities and flexible working arrangements, several studies also highlight precarious working conditions, job instability, and limited protection for workers on digital platforms [51]. As in the previous paragraph (2.1), to summarise the main effects of the digital transition on human well-being identified in the literature, Table 2 is presented, which highlights the positive and negative aspects of each analysed theme.

Table 2. The effects of digital transition on human well-being

TOPIC	POSITIVE EFFECTS	NEGATIVE EFFECTS
Digital work (smart working, platforms, gig economy))	• Greater flexibility in working arrangements; • Reduction of travel; • New opportunities in digital work	• Blurred boundaries between work and private life; • Stress and burnout; • Professional isolation; • Job insecurity
Intensive use of digital technologies	• Access to information and immediate communication; • Social connectivity;	• Technostress; • Information overload; • Mental fatigue
Digital consumption and online platform	• Time savings in shopping; • Accessibility to goods and services	• Compulsive consumption; • Acceleration of the replacement cycles of goods; • Stress and psychological discomfort
Digital access and social inclusion	• Greater social integration; • Access to services and information; • Increase in subjective well-being	• Digital divide
Access to digital education and training	• Democratisation of access to education; • Better results	• Differences in digital skills

Overall, the literature highlights how the impacts of the digital transition on human well-being are deeply ambivalent: alongside benefits in terms of access, flexibility, and inclusion, there are critical issues related to stress, isolation, surveillance, and digital inequalities, confirming the need for governance capable of guiding digitalisation towards socially sustainable outcomes.

3 Methodological Approaches to Managing the Digital Transition: Hard-Normative and Soft-Cultural Tools

As highlighted by the literature in paragraph 2, the shift of face-to-face activities towards digital forms not only reduces environmental impact but also involves a complex reorganisation of new material and energy flows, through which the new global technological infrastructure is activated, with possible negative environmental effects. From the perspective of daily practices, a significant transformation is observed: digitalisation offers numerous advantages, such as flexible working (smart and remote work), new opportunities to access services, information, training, and socialisation. However, it also presents negative effects on psychological well-being: many forms of digital stress, isolation, and inequality are indeed evident. Digitalisation, therefore, cannot be considered inherently sustainable or unsustainable: its impact largely depends on the institutional, cultural, and energy framework within which it is implemented. In this context, the main challenge lies in conscious management of the digital transition, that is, in the ability to reduce or at least mitigate negative impacts, both from an environmental perspective and in terms of human well-being. Analysis of the literature reveals the need for a systemic approach that integrates regulatory tools (hard) and cultural (soft) factors, as well as multi-level governance, to reconcile technological progress, environmental sustainability, and human well-being, and to counteract the negative effects of digitalisation. This work, therefore, proposes a definition of these regulatory tools to assist institutions in guiding the digital transition towards objectives of environmental sustainability and human well-being.

3.1 Hard-Normative Tools

From a regulatory perspective, several studies highlight the need to strengthen environmental regulation in the ICT sector. Energy efficiency standards for data centres and digital infrastructure, transparency obligations regarding emissions, and incentives for the use of renewable energy sources are essential tools to reduce the climate impact of digitalisation. Some authors also propose introducing mandatory environmental impact assessments for large digital projects, like those required for other industrial infrastructure projects. Another area of intervention concerns the sustainable design of digital technologies. From this perspective, the principle of ‘sustainability by design’ should be adopted as a genuine criterion for evaluation within public regulation. This implies that technology companies and policymakers systematically incorporate, from the early design stages, a checklist of key questions related to the entire lifecycle of the digital product, capable of translating negative effects into positive outcomes of the digital transition. These checklists can serve as a powerful operational tool to support public administration, enabling systematic verification and monitoring of the environmental

and social sustainability of digital solutions before implementation. An example of such a checklist, addressing the negative effects on the environment, could be as follows (Table 3):

Table 3. Example of an operational checklist

TOPIC	ANSWERS	NEGATIVE EFFECTS	POSSIBLE ACTIONS
Energy efficiency	1. Is the planned infrastructure truly necessary? 2. Are minimum energy efficiency standards expected? 3. Does the energy used come from renewable sources?	• Increase in the necessary digital infrastructure; • Increase in computational consumption	• Standard energy practices for data centres; • Use of renewables; • Load optimisation; • Assessment of energy impact
Dematerialization	4. Does the digital solution truly replace a physical function or does it add to it? 5. Have the environmental benefits been measured across the entire lifecycle?	• Rebound effect; • Increase in digital consumption	• LCA assessment; • Monitoring of uses; • Limits on the duplication of services and processes
Circular Economy and traceability	6. Does the technology used truly improve traceability? 7. Does the environmental benefit outweigh the cost of infrastructure? 8. Do the collected data lead to concrete actions of reuse and recycling?	• Technological complexity; • Increased use of digital infrastructures	• Minimum traceability requirements; • Data interoperability; • Circular design standards

(*continued*)

Table 3. (*continued*)

TOPIC	ANSWERS	NEGATIVE EFFECTS	POSSIBLE ACTIONS
Smart cities	9. Are the connected devices proportionate to the expected benefits? 10. Is an impact assessment on energy, water, and soil planned? 11. Is the territory capable of absorbing the costs?	• Increase in energy demand; • Increase in connected device	• Locational planning; • Sustainability thresholds; • Preventive environmental assessments; • Territorial governance
AI and digital infrastructureale	12. Are there less energy-intensive alternatives? 13. Are there measures in place for efficiency and for managing the hardware's lifecycle?	• High energy consumption; • • Use of data centres; • Growth of e-waste	• Criteria of necessity for use; • Efficiency standards; • Sustainable procurement

This checklist can be operationally adopted by public decision-makers as an evaluative toolkit to guide calls for proposals, incentives, and authorisations, and constitutes a structural element of hard-normative instruments. By making 'sustainability by design" a systematic and mandatory criterion, public policies gain a new tool to steer digital transformation towards tangible environmental sustainability objectives.

3.2 Soft-Cultural Solutions and Digital Literacy

Alongside regulatory tools, literature emphasises the importance of cultural and educational interventions. The spread of digital technologies indeed makes a new form of literacy necessary, aimed at promoting a conscious and sustainable use of technological tools. It is for this reason that, as with nutritional or environmental education, this new form of digital education should also be integrated to develop critical skills in the use of digital media, which can help prevent phenomena such as technological dependence, information overload, and digital stress. In this context, there is also an increasing debate on the right to disconnect, considered by many scholars as a fundamental tool to protect workers' psychological well-being in the era of hyperconnectivity and to establish clearer boundaries between professional and private life.

3.3 Hard and Soft Integrations for Regional Planning

The Italian urban planning policy, based on Law 1150/42, is outdated and no longer responds to the dynamics of contemporary society. In this context, it is evident that there is a need to move beyond compartmentalised sectoral policies and to promote a planning approach with multiple interpretative keys, capable of integrating digital, climate, and social policies. The effectiveness of the digital transition indeed depends on the ability to coordinate technological innovation strategies with the objectives of decarbonisation, circular economy, and the protection of human well-being.

The governance of the territory and urban planning policies must therefore be restructured in both their methodological aspects and their analysis and forecasting. As extensively discussed in the previous paragraphs, the use of digital technologies and remote work profoundly alters mobility patterns, migratory flows, and settlement dynamics, necessitating the update of planning tools traditionally based on historical data. A concrete example of this evolution is the use of real-time urban dashboards that collect and display data from sensors distributed throughout the city (traffic, noise, air quality, pedestrian flows). These platforms allow urban planners to monitor infrastructure usage in real time and anticipate overload or emerging needs. For instance, detecting significant increases in traffic flows during certain time slots through analysis of aggregated data can support targeted decisions, such as temporary modifications to traffic management or the regulation of access in ZTLs (Limited Traffic Zones), or dynamically guide public transport planning. Similarly, a growing trend of increased use of public spaces in peripheral areas, revealed by heatmaps generated from urban monitoring systems, could suggest revising zoning criteria or allocating new resources for services and urban greenery in those districts, translating digital analysis into concrete actions for territorial governance. To fully understand these phenomena, planning activities must be as interdisciplinary as possible, integrating analyses of the lifecycle of digital infrastructures, longitudinal studies on the effects of technology use on psychological well-being, and research that combines environmental, social, and economic indicators. Ultimately, digital transition can only be considered truly sustainable if it is managed as an integrated process, capable of maximising technological benefits while minimising environmental and social costs. Only through a coordinated approach involving technological innovation, public policies, and cultural transformations will it be possible to steer digitalisation towards development models that are compatible with the planet's ecological limits and the improvement of human quality of life. Considering these considerations, the following paragraph proposes an operational framework to translate the effects of the digital transition into actions and tools for good land governance.

4 Operational Proposal for Territorial Planning and Programming

Digital transition, as highlighted in the literature review, does not produce linear or unambiguous effects: it can reduce certain environmental pressures (e.g., commuting mobility) but also generate new ones (energy consumption, logistics, ICT infrastructure); it can expand opportunities for access to services and employment, but also exacerbate inequalities, isolation, and vulnerabilities (digital divide, precariousness, digital stress). However, negative effects should not be viewed solely as elements to be mitigated, but

rather as territorial and social signals that, if properly identified, can guide the development of governance tools capable of steering towards a more sustainable and balanced digitalisation. In this context, the built environment, and in particular residential patterns, becomes a crucial area for observation and intervention: the increasing flexibility of work times and locations (smart working, hybrid functions, online services) tends to recalibrate housing choice criteria and redefine models of territorial attractiveness. This creates a need for Public Administrations and territorial governance bodies to equip themselves with knowledge and assessment tools capable of capturing these transformations in an integrated manner, anticipating trade-offs and guiding policies towards sustainability, inclusion, and quality of life.

Building on these considerations, this contribution proposes an operational framework to interpret the effects of the digital transition through the lens of residential well-being and to translate them into concrete territorial governance actions. The procedure, tested in the PRIN 2022 PNRR Research, is based on the integration of three dimensions of residential well-being: (i) territorial-objective dimension, (ii) economic dimension, and (iii) subjective dimension. The fundamental assumption is that none of these components, considered in isolation, is sufficient to guide robust decisions: territorial analysis, derived from planning tools, provides the material and functional structure of contexts; market analysis explicates the selective mechanisms regulating access to residence; and the analysis of preferences and perceptions allows understanding the emerging housing demand, which today is also influenced by the digitalisation of daily practices. The integration of the three dimensions, however, enables the construction of unified interpretative frameworks useful for reading convergences and misalignments (for example, areas with high urban quality but economically inaccessible; accessible areas but perceived as lacking identity; territories with high environmental quality but with deficits in services and connections), transforming them into governance scenarios and operational guidelines. Operationally, the protocol can be organised into a sequence of modular and scalable phases tailored to the intervention context, each culminating in outputs immediately usable by the Public Administration: maps, synthetic indicators, and priority matrices. After each phase, explicit decision points (go/no-go checkpoints) are introduced, allowing the administration to evaluate the results obtained and decide transparently whether to proceed, deepen the analysis, or reallocate resources. This approach enhances the clarity and practicality of the protocol, facilitating operational decisions throughout the process. Below are the phases of the operational framework, subsequently summarised in a graphical form (Fig. 1):

1. **Construction of the territorial knowledge framework** (*objective dimension*). The first component concerns the reconstruction of urban and territorial quality through objective indicators, consistent with the strategies and contents of planning tools (territorial/metropolitan plans, structural/operational plans, PUMS, sectoral and civil protection plans). The aim is not to produce a mere inventory but to structure a set of criteria capable of representing, at an appropriate scale, the conditions of liveability and accessibility: service provisions, infrastructure, environmental quality, risks, connectivity and digital provisions, quality of the building heritage. In this phase, the public administration can enhance official open sources and institutional databases to ensure transparency, up-to-date information, and replicability. Processing in a GIS

environment enables the creation of comparable cartographic layers and the identification of territorial gradients and discontinuities, which are useful for guiding rebalancing policies.

2. **Analysis of economic accessibility and the residential market** (*economic dimension*). The second component aims to measure the sustainability of living and the socio-spatial selection mechanisms induced by the market. The analysis integrates homogeneous micro-areas of the real estate market (zones and subzones), property values (OMI and/or complementary sources), and income distribution to calculate affordability indicators (e.g., income-to-value ratios or indices based on sustainability thresholds). This dimension is essential to 'make visible' the gap between territorial quality and actual accessibility: in highly attractive contexts, digitalisation can further increase settlement pressure and competition for access, amplifying phenomena of gentrification and displacement. For public administration, these results provide a technical basis for regulatory and rebalancing policies (social housing, subsidised housing, equalisation tools, public-private agreements, rent containment measures), distinguishing between areas where the criticality is mainly 'market-driven' and those where it is instead 'related to infrastructure'.
3. **Analysis of preferences, perceptions, and propensities for mobility** (*subjective dimension*). The third component examines subjective well-being and housing demand using primary data (surveys) or participatory tools to assess residential satisfaction, perceived quality, emerging needs, and willingness to relocate. In a digital transition context, this level becomes crucial because it captures transformations that cannot be deduced solely from structural data: for example, an increased demand for domestic spaces suitable for hybrid work, greater attention to green spaces and environmental quality, or new preferences related to proximity to digital services and fast connections. The results can be categorised by territorial type (city centres, intermediate hubs, peri-urban and mountain areas), assigning synthetic scores to areas and producing maps of 'demand for well-being', useful for understanding where the digital transition is effectively opening new settlement trajectories.
4. **Integration and multi-criteria reading for governance** (*decision support tools*). The three dimensions are then integrated using SMCDA procedures, selecting techniques consistent with the administrative objective: when it is necessary to produce a synthetic index and a comparable spatial classification, a non-linear compensatory approach (e.g., MAUT) proves effective, capable of modelling thresholds and non-proportional effects, while maintaining transparency and reproducibility. The integration is not only for descriptive purposes: it allows the construction of scenario matrices that associate recurring performance combinations (high/low for each dimension) with typical criticalities and policy packages. In this way, the final output is not just a map, but a set of 'governance classes' (territories to regulate, territories to enhance, territories to rebalance, territories to make accessible), directly translatable into strategies and tools.
5. **From knowledge to action: hard and soft approaches**. Based on the integrated results, the Public Administration can structure actions into two complementary plans. At this stage, the operational checklists proposed in the work serve as decision support, enabling the translation of issues identified through the multidimensional analysis of residential well-being into evaluation criteria and policy directions.

On the hard-normative level, interventions include measures on planning tools and programmes: guidelines for the regeneration of existing assets (rather than expansion), standards and provisions, housing policies and market regulation, sustainable infrastructure and mobility, and criteria for the location of digital infrastructures with impact assessments (energy, water, land). On the soft-cultural level, digital literacy and inclusion (reducing the digital divide), the promotion of community practices and place-making, participatory tools, and public communication aimed at raising awareness of digital impacts (digital hygiene, right to disconnect, work-life balance) are central. The core idea is that a credible "twin transition' requires technological innovation to be accompanied by territorial policies oriented towards social and environmental justice, preventing the benefits of digitalisation from being concentrated in a few areas while costs are borne elsewhere.

In summary, the proposal outlines an operational decision-support device for territorial governance bodies, based on replicable protocols, transparent datasets, and cartographic and decision-making outputs. Its utility lies in its ability to interpret the transformations induced by the digital transition not as abstract phenomena, but as territorialised, differentiated, and governable processes. The integration of plan analysis, market reading, and subjective demand interpretation allows strategic planning and, subsequently, operational programming to be directed towards selective and coherent interventions, aimed at maximising the benefits of digitalisation while minimising environmental and social costs, with particular attention to residential living and quality of life. However, the framework may present some methodological limitations, including, for example, the variability and reliability of territorial and economic data, which can vary significantly between different urban contexts, affecting the comparability of outcomes, as well as biases in the detection of subjective preferences, which, although providing valuable information, can vary depending on the selection or representativeness of the analysed samples.

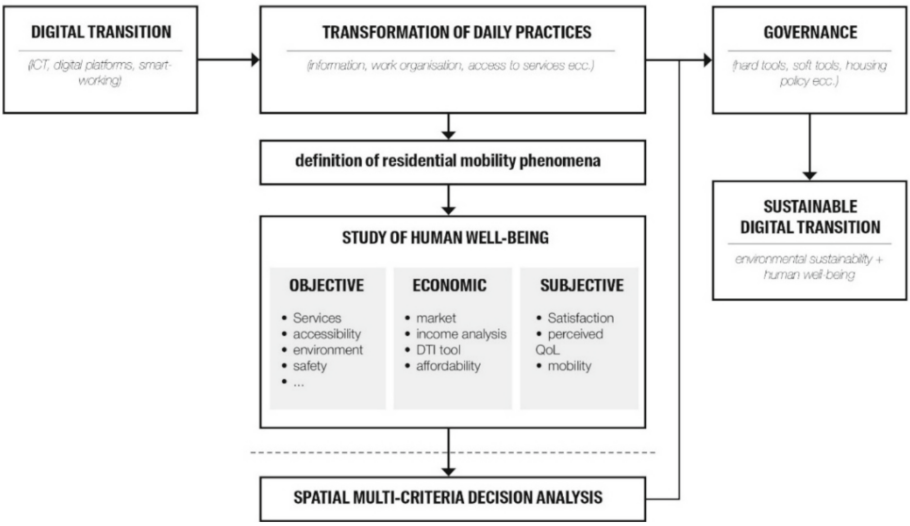


Fig. 1. Conceptual framework of digital transition, residential well-being and territorial governance

5 Conclusions

Digital transformation is one of the most important processes in modern societies, but its effects are ambivalent, creating opportunities and critical issues both environmentally and socially. From an environmental perspective, it can reduce some pressures through dematerialisation and energy optimisation, but it also increases energy, water, and critical resource consumption, as well as the production of electronic waste. Recent literature emphasises how digitalisation transforms, rather than eliminates, environmental impacts, making them less visible and more distributed across space. This questions its environmental sustainability, requiring a systemic approach that considers the entire lifecycle of digital technologies, from materials to waste, including induced effects on consumption patterns. Socially, digitalisation alters daily practices and relationships: it increases access and flexibility, but also generates vulnerabilities such as technostress, isolation, and inequalities, thereby widening the digital divide.

In this perspective, the contribution of this work lies in highlighting the need for analytical and operational tools capable of reading the environmental and social effects of the digital transition in an integrated manner. In particular, the methodological proposal outlined in the previous paragraph suggests adopting multidimensional approaches that combine territorial, economic, and subjective analyses of residential well-being to support decision-making by public administrations and territorial governance bodies. The integration of territorial data, market dynamics, and residents' perceptions allows for the identification of misalignments and emerging critical issues in settlement processes, providing a useful knowledge base to guide planning and programming policies that are more aligned with the objectives of the so-called twin transition, namely the integration of digital transition and ecological transition. Within this analytical framework, the work

also proposes the use of operational decision-support tools, such as assessment checklists, to translate the negative effects of the digital transition identified in the literature into analysis criteria and guiding questions to shape public policies. Checklists are a very useful operational tool, capable of linking the environmental and social criticalities of the digital transition to the three dimensions of residential well-being, supporting public decision-makers in defining targeted actions. In conclusion, the digital transition can truly contribute to improving the quality of life and reducing pressure on ecosystems only if it is managed as a multidimensional process, capable of integrating technological innovation, environmental sustainability, and social justice. In the absence of such an integrated approach, there is a risk that the benefits of digitalisation will be concentrated in certain exclusive areas or territories, while environmental and social costs are transferred elsewhere, making digital transformation less equitable and less sustainable in the long term.

References

1. Batty, M.: Big data, smart cities and city planning. Dialogues Hum. Geogr. **3**(3), 274–279 (2013). https://doi.org/10.1177/2043820613513390
2. Kitchin, R.: The real-time city? Big data and smart urbanism. GeoJournal **79**(1), 1–14 (2014). https://doi.org/10.1007/s10708-013-9516-8
3. IPCC: Climate Change 2023: Synthesis Report. Intergovernmental Panel on Climate Change (2023). https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf. Accessed 4 Mar 2026
4. WHO: World Mental Health Report: transforming mental health for all. World Health Organization (2022). https://www.who.int/publications/i/item/9789240049338. Accessed 4 Mar 2026
5. Lange, S., Pohl, J., Santarius, T.: Digitalization and energy consumption: does ICT reduce energy demand? Ecol. Econ. **176**, 106760 (2020). https://doi.org/10.1016/j.ecolecon.2020.106760
6. Oliveira, T.C., Barlow, J., Gonçalves, L., Bayer, S.: Teleconsultations reduce greenhouse gas emissions. J. Health Serv. Res. Policy **18**(4), 209–214 (2013). https://doi.org/10.1177/1355819613492717
7. Holmner, Å., Ebi, K.L., Lazuardi, L., Nilsson, M.: Carbon footprint of telemedicine solutions - unexplored opportunity for reducing carbon emissions in the health sector. PLoS ONE **9**(9), e105040 (2014). https://doi.org/10.1371/journal.pone.0105040
8. Li, W., Liu, N., Long, Y.: Assessing carbon reduction benefits of teleworking: a case study of Beijing. Sci. Total. Environ. **889**, 164262 (2023). https://doi.org/10.1016/j.scitotenv.2023.164262
9. Kitou, E., Horvath, A.: Energy-related emissions from telework. Environ. Sci. Technol. **37**(16), 3467–3475 (2003). https://doi.org/10.1021/es025849p
10. Mazmanian, M., Orlikowski, W.J., Yates, J.: The autonomy paradox: the implications of mobile email devices for knowledge professionals. Organ. Sci. **24**(5), 1337–1357 (2013). https://doi.org/10.1287/orsc.1120.0806
11. Shi, Y., Sorrell, S., Foxon, T.: The impact of teleworking on domestic energy use and carbon emissions: an assessment for England. Energy Build. **287**, 112996 (2023). https://doi.org/10.1016/j.enbuild.2023.112996
12. Viu-Roig, M., Alvarez-Palau, E.J.: The impact of e-commerce-related last-mile logistics on cities: a systematic literature review. Sustainability **12**(16), 6492 (2020). https://doi.org/10.3390/su12166492

13. Dempsey, N., Bramley, G., Power, S., Brown, C.: The social dimension of sustainable development: defining urban social sustainability. Sustain. Dev. **19**(5), 289–300 (2011). https://doi.org/10.1002/sd.417
14. Morozov, E.: To save everything, click here: the folly of technological solutionism. PublicAffairs (2013)
15. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press (2021)
16. Pötzsch, H.: Materialist perspectives on digital technologies: informing debates on digital literacy and competence. Nordicom Rev. **37**(1), 77–95 (2016). https://doi.org/10.1515/nor-2016-0006
17. Gajarawala, S.N., Pelkowski, J.N.: Telehealth benefits and barriers. J. Nurse Pract. **17**(2), 218–221 (2021). https://doi.org/10.1016/j.nurpra.2020.09.013
18. Latupeirissa, J.J.P., Dewi, N.L.Y., Prayana, I.K.R., Srikandi, M.B., Ramadiansyah, S.A., Pramana, I.B.G.A.Y.: Transforming public service delivery: a comprehensive review of digitization initiatives. Sustainability **16**(7), 2818 (2024). https://doi.org/10.3390/su16072818
19. Lyzwinski, L.N.:. Organizational and occupational health issues with working remotely during the pandemic: a scoping review of remote work and health. J. Occup. Health **66**(1), uiae005 (2024). https://doi.org/10.1093/joccuh/uiae005
20. Wells, J., et al.: A systematic review of the impact of remote working referenced to the concept of work–life flow on physical and psychological health. Workplace Health Saf. **71**(11), 507–521 (2023). https://doi.org/10.1177/21650799231176397
21. Hilty, L.M., Aebischer, B.: ICT for sustainability: an emerging research field. In: Hilty, L., Aebischer, B. (eds.) ICT Innovations for Sustainability. Advances in Intelligent Systems and Computing, vol. 310, pp. 3–36. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09228-7_1
22. European Commission: The twin green and digital transition: Key requirements for successful implementation. European Commission (2022)
23. International Energy Agency (IEA): Energy efficiency and digitalisation (2019)
24. Cagno, E., et al.: Energy management and Industry 4.0: Analysis of the interactions and future research directions. Energy (2025)
25. Lu, Y.: Industry 4.0: A survey on technologies, applications and open research issues. J. Ind. Inf. Integr. **6**, 1 (2017). https://doi.org/10.1016/j.jii.2017.04.005
26. Charfeddine, L., Umali, M.: ICT sector, digitization and environmental sustainability: a systematic review of the literature from 2000 to 2022. Renew. Sustain. Energy Rev. **184**, 113482 (2023). https://doi.org/10.1016/j.rser.2023.113482
27. Coroama, V.C., Moberg, Å., Hilty, L.M.: Dematerialization through electronic media?. In: Hilty, L., Aebischer, B. (eds.) ICT Innovations for Sustainability. Advances in Intelligent Systems and Computing, vol. 310, pp. 405–421. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09228-7_24
28. Saberi, S., Kouhizadeh, M., Sarkis, J., Shen, L.: Blockchain technology and its relationships to sustainable supply chain management. Int. J. Prod. Res. **57**(7), 2117–2135 (2019). https://doi.org/10.1080/00207543.2018.1533261
29. Kouhizadeh, M., Sarkis, J.: Blockchain practices, potentials, and perspectives in greening supply chains. Sustainability **10**(10), 3652 (2018). https://doi.org/10.3390/su10103652
30. International Energy Agency: Energy and AI: Energy demand from AI. IEA (2025)
31. Belkhir, L., Elmeligi, A.: Assessing ICT global emissions footprint: trends to 2040. J. Clean. Prod. **177**, 448–463 (2018)
32. Jones, N.: How to stop data centres from gobbling up the world's electricity. Nature **561**, 163–166 (2018)
33. Lange, S., et al.: Digitalization and energy consumption. Environ. Innov. Societal Transitions (2020)

34. Sovacool, B.K., Hook, A., Martiskainen, M., Brock, A., Turnheim, B.: The decarbonisation divide: contextualizing landscapes of low-carbon exploitation and toxicity in Africa. Glob. Environ. Chang. **60**, 102028 (2020). https://doi.org/10.1016/j.gloenvcha.2019.102028
35. He, P., Li, J., Shen, L., Tao, F., Ju, Y., Wang, J.: Toward green smart cities using Internet of Things and artificial intelligence: a review. Sustain. Comput. Inform. Syst. **36**, 100822 (2022). https://doi.org/10.1016/j.suscom.2022.100822
36. Lagarde, C., Robillart, M., Bigaud, D., Pannier, M.L.: Assessing and comparing the environmental impact of smart residential buildings: a life cycle approach with uncertainty analysis. J. Clean. Prod. **467**, 143004 (2024)
37. Boccoli, G., Sestino, A., Gastaldi, L., Corso, M.: The impact of autonomy and temporal flexibility on individuals' psychological well-being in remote settings. Sinergie – Ital. J. Manage. **40**(2), 327–349 (2022). https://doi.org/10.7433/s118.2022.15
38. Wells, J., et al.: A systematic review of the impact of remote working referenced to the concept of work-life flow on physical and psychological health. Workplace Health Saf. **71**(11), 507–521 (2023). https://doi.org/10.1177/21650799231176397
39. Deschênes, A.A.: Professional isolation and pandemic teleworkers' satisfaction and commitment: the role of perceived organizational and supervisor support. Eur. Rev. Appl. Psychol. **73**(2), 100823 (2023). https://doi.org/10.1016/j.erap.2022.100823
40. Kumar, P.S.: TECHNOSTRESS: a comprehensive literature review on dimensions, impacts, and management strategies. Comput. Hum. Behav. Rep. **16**, 100475 (2024). https://doi.org/10.1016/j.chbr.2024.100475
41. Tafesse, W., Aguilar, M.P., Sayed, S., Tariq, U.: Sovraccarico digitale, meccanismi di coping e coinvolgimento degli studenti: un'indagine empirica basata sul framework SOR. Sage Open **14**(1) (2024)
42. Twenge, J.M., Campbell, W.K.: Media use and mental health: a review. Curr. Opin. Psychol. (2019)
43. Twenge, J.M., et al.: Increases in depressive symptoms among U.S. adolescents after 2010. Clin. Psychol. Sci. (2018)
44. Primack, B.A., et al.: Social media use and perceived social isolation among young adults in the U.S. Am. J. Prev. Med. (2017)
45. Przybylski, A.K., et al.: Motivational, emotional, and behavioral correlates of fear of missing out (FOMO). Comput. Hum. Behav. (2013)
46. Black, D.W.: Compulsive shopping: a review and update. Curr. Opin. Psychol. **46**, 101321 (2022). https://doi.org/10.1016/j.copsyc.2022.101321
47. Kearns, A., Whitley, E.: Associations of internet access with social integration, wellbeing and physical activity among adults in deprived communities: Evidence from a household survey. BMC Public Health **19**, 860 (2019). https://doi.org/10.1186/s12889-019-7199-x
48. Ragnedda, M., Muschert, G.W. (eds.) The Digital Divide: The Internet and Social Inequality in International Perspective, 1st edn. Routledge (2013). https://doi.org/10.4324/9780203069769
49. Veletsianos, G., Shepherdson, P.: A systematic analysis and synthesis of the empirical MOOC literature published in 2013–2015. Int. Rev. Res. Open Distrib. Learn. **17**(2), 198–221 (2016). https://doi.org/10.19173/irrodl.v17i2.2448
50. Means, B., Toyama, Y., Murphy, R., Bakia, M.: The effectiveness of online and blended learning: a meta-analysis of the empirical literature. Teach. Coll. Rec. **115**(3), 1–47 (2013). https://doi.org/10.1177/016146811311500307
51. De Stefano, V.: The rise of the "just-in-time workforce": on-demand work, crowdwork and labour protection in the gig economy. Comp. Labor Law J. **37**(3):471–504 (2016). A Publication of the U.S. National Branch of the International Society for Labor Law and Social Security [and] the Wharton School, and the Law School of the University of Pennsylvania

Evaluating Alternative Continuous Metaheuristic Solution Encodings for Routing Problems with Multiple Drones with Interceptions

Sarah Dillon and Jacomine Grobler(✉)

Department of Industrial Engineering, Stellenbosch University, Stellenbosch, South Africa
sarahdillon999@gmail.com, jacomine.grobler@gmail.com

Abstract. The integration of trucks and drones has emerged as a promising approach for improving the efficiency of last-mile delivery operations. This paper studies the travelling salesman problem with multiple drones and interception, in which a truck carries multiple drones that may be launched from the truck and retrieved while the truck is in motion. Recent work has shown that continuous metaheuristic algorithms are effective for this problem class; however, the impact of solution encoding on computational performance has not yet been systematically examined. Two alternative solution encodings are compared in this paper: an assignment strategy that explicitly allocates each customer to a specific vehicle, and a more general encoding that assigns customers to either the truck or the next available drone without prespecifying which drone. Both encodings are embedded within identical continuous metaheuristic frameworks and evaluated using two algorithms across 15 benchmark instances with varying spatial distributions and sizes. Computational results demonstrate that the more general encoding consistently achieves substantial reductions in runtime while maintaining comparable solution quality. The findings highlight the critical role of encoding design in continuous optimisation for hybrid vehicle routing problems.

Keywords: Drone delivery · TSP with multiple drones with interceptions · Evolutionary algorithms · Swarm intelligence · Last-mile

1 Introduction

The optimisation of last mile delivery logistics has become increasingly relevant due to the rapid growth of e-commerce in recent years. E-commerce, otherwise known as online shopping, refers to the use of the internet for commercial transactions [1]. With widespread internet access and the growth of smartphone and computer usage, retail companies leverage e-commerce to reach consumers more

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 538–555, 2027.
https://doi.org/10.1007/978-3-032-30494-0_34

effectively. The surge in online shopping has led to an increased volume in the last mile, and this surge is accompanied by heightened customer expectations. These pressures have intensified interest in delivery systems that make use of drones, or *unmanned aerial vehicles* (UAV), in order to improve operational efficiency beyond what is achievable with truck-only fleets.

Drones offer high travel speeds and the ability to bypass road congestion; however, their limited payload capacity and flight endurance prevent them from fully replacing ground vehicles [2]. Consequently, research has increasingly focused on integrated truckâĂŞdrone systems, which have been shown to reduce delivery times and operational costs when compared to conventional truck-based routing [3]. The growing practical relevance of such systems is reflected in recent industry deployments. Since the first commercial autonomous drone delivery in 2016 [4], adoption has accelerated steadily, with thousands of deliveries performed daily by 2022 and hundreds of thousands annually by 2023 [5,6]. Major logistics providers, including Amazon, Wing, Flytrex, Alibaba, JD Logistics, and Walmart, have incorporated drone technology into their delivery strategies [7].

From an optimisation perspective, many coordinated truckâĂŞdrone delivery problems have been formalised as extensions of the *travelling salesman problem with a drone* (TSPD). The original TSPD was introduced by Agatz *et al.* [8], and extended to allow drones to intercept the moving truck between customer locations [9]. Allowing for multiple drones onboard the truck to intercept the truck en route yields the *travelling salesman problem with multiple drones and interception* (TSPmDi) [10].

Existing continuous metaheuristic approaches to the TSPmDi rely on a solution encoding in which each customer is explicitly assigned either to the truck or to a specific drone [10]. This encoding, referred to as *Solution Encoding 1* (SE1), requires repeated discretisation across all drones and customers, introducing significant computational overhead. This paper investigates an alternative modelling approach, termed *Solution Encoding 2* (SE2), in which customers are assigned either to the truck or to the next available drone, without pre-selecting a specific drone. The impact of this encoding on algorithmic performance is evaluated using 15 benchmark datasets, with varying sizes and spatial distributions. A *self adaptive neighbourhood search differential evolution* (SaNSDE) algorithm and a *guaranteed convergence particle swarm optimisation* (GCPSO) algorithm, both combined with a *nearest neighbour heuristic for initial solutions* (NNHis) operator, is applied to the TSPmDi for both SE1 and SE2. By modifying only the assignment encoding, the study isolates the performance and computational effect of encoding design.

To the authors' knowledge, this is the first study to explicitly quantify how solution encoding influences the performance and computational efficiency of continuous metaheuristics for the TSPmDi. This paper demonstrates that encoding-level design choices can dominate runtime performance without affecting solution quality.

The paper is organised as follows: Sect. 2 reviews related work, and Sect. 3 outlines the metaheuristic algorithms used in this paper. The problem is

described in Sect. 4. The proposed continuous metaheuristic algorithm is introduced in Sect. 5. The empirical evaluation of the metaheuristic algorithms is documented in Sect. 6. A conclusion of the paper is presented in Sect. 7.

2 Literature Review

Recent research has increasingly focused on cooperative truck-drone systems for last-mile delivery. Several comprehensive survey papers review developments in this area, covering modelling variants, solution approaches, and emerging applications [7,11–17].

Two well-studied single-truck and single-drone variants are the *travelling salesman problem with a drone* (TSPD) [8] and the *flying sidekick travelling salesman problem* (FSTSP) [18,19]. Moremi [9] extended the TSPD to define the *TSPD with interception* (TSPDi), which allows the drone to intercept the truck while the truck is in motion between customer nodes. An *ant colony system* (ACS) algorithm was developed to solve this problem. Optimal solutions for the TSPDi were later obtained by Knoetze [20], although the computational time limit of eight hours resulted in no optimal solutions for instances containing 50 or more nodes. Ernst [21] subsequently proposed continuous metaheuristic algorithms for the TSPDi, which outperformed both the ACS approach and the exact method, indicating the significant potential of continuous metaheuristics for this problem.

Most *TSP with multiple drones* (TSPmD) problems in the literature require the truck and drones to rendezvous at customer locations after completing their respective deliveries with the first-arriving vehicle waiting for the other. However, increasing research is being conducted on models that allow drones to intercept the truck at non-customer locations. Table 1 summarises existing interception-based systems that consider multiple drones.

These studies can be broadly divided into two categories. The first assumes predefined launch and rendezvous locations. Betti *et al.* [22,23] allowed each customer location to be associated with corresponding launch and rendezvous points that may replace customer nodes for drone operations, while Salama [24] and Han *et al.* [25] considered a variant with predefined truck stop locations from which drones may be launched or retrieved. The second category permits drone launches and retrievals anywhere along the truck's route between two customer nodes. Within this group, some studies restrict the truck to a road network. Lin *et al.* [26,27] and Narayan [28] adopted this assumption. Other studies removed the road network restriction entirely [10,29,30]. The Euclidean distance is the most common distance metric used in the literature [10,23–27,30].

No existing work investigates alternative solution encodings for continuous metaheuristics or investigates the effect of adding more drones to the system when solving the TSPmDi with continuous metaheuristics. In particular, no existing work evaluates how different continuous solution encodings affect feasibility enforcement, repair behaviour, and runtime efficiency, when solving the TSPmDi.

Table 1. Summary of relevant work on truck and multiple drone interception systems

	Reference	Drones	Nodes	Method	Objective
Predefined locations	Betti *et al.* [22]	1, 3, 5	25–100	Greedy heuristics	Maximise reward
	Betti *et al.* [23]	1, 3, 5	25–100	Knapsack approximation	Minimise time
	Han *et al.* [25]	2	10–50	Variable Neighbourhood Search (VNS)	Minimise waiting
	Salama *et al.* [24]	4	8–59	Simulated annealing, VNS, rule-based heuristics	Minimise time
Interception Allowed	Lin *et al.* [27]	2	10–100	Genetic Algorithm (GA)	Minimise time
	Lin *et al.* [26]	1–5	20–100	Hybrid GA/PSO	Minimise time
	Narayan [28]	1–11	28	Exact methods	Minimise time
	Li [29]	2	5–30	Adaptive Large Neighbourhood Search (ALNS)	Minimise cost
	Thomas [30]	3, 6	10–100	Heuristics	Minimise time
	Dillon *et al.* [10]	2	10–250	Continuous metaheuristics	Minimise time
	This Paper	**2–4**	**10–250**	**Continuous metaheuristics**	**Minimise time**

3 Algorithm Background

The SaNSDE and GCPSO algorithms are explained in this section, along with the *nearest neighbour heuristic for initial solutions* (NNHis) initialisation strategy.

3.1 Self Adaptive Neighbourhood Search Differential Evolution

Differential evolution (DE) is a population-based, stochastic optimisation algorithm that relies on difference vectors to determine step sizes between individuals in the population. These differences drive the search toward improved solutions. The difference mutation operator enables step sizes to adapt automatically to the objective function landscape [31]. In the early stages of the search, when individuals are widely dispersed across the search space, larger step sizes promote exploration and help avoid local optima. As the population converges, smaller step sizes emerge, allowing for refined exploitation around promising regions. The SaNSDE algorithm applied in this paper was introduced by Yang *et al.* [32]. The algorithm combines the best features of the *Neighbourhood Search DE* (NSDE) [32] and the *Self-adaptive DE* (SaDE) [33] algorithms.

3.2 Guaranteed Convergence Particle Swarm Optimisation

Particle swarm optimisation (PSO) is a swarm intelligence optimisation algorithm developed by Kennedy and Eberhart [34], inspired by the collective move-

ment of bird flocks. The algorithm iteratively improves candidate solutions by moving particles through a multi-dimensional hyperspace, where each particle's position represents a candidate solution. Velocity and displacement updates are applied to guide particles to a new position, thereby obtaining new solutions in the search space [35]. At any time step t, a particle's velocity is determined by its previous velocity, its personal best position, and the global best position found by the swarm. Convergence occurs as particles collectively move towards these best-known positions.

To prevent stagnation, the *guaranteed convergence PSO* (GCPSO) introduces a random search by the global best particle around the global best particle position [36]. The area of the search adapts according to the number of consecutive successful or unsuccessful attempts to improve the objective function value.

3.3 The Nearest Neighbour Heuristic for Initial Solutions

The NNHis initialisation strategy introduces six high-quality initial solutions to the initial population based on nearest-neighbour construction using truck-only and alternating truck-drone assignments [21,37]. The NNHis operator is adapted from [21] for the TSPmDi by introducing the following six initial solutions into the initial population of the algorithm:

- The first solution is generated using a *nearest neighbour heuristic* (NNH). The depot is considered the first node. The node with the smallest Euclidean distance from the depot is chosen as the second node to visit. Thereafter, the node closest to the second node is allocated as the next stop. This process is continued until all nodes have been allocated and the final node is again assigned as the depot. All nodes are to be visited by the truck, and the drone is stationary.
- The second solution is generated using two parallel NNHs. Two different routes are generated, both starting from the depot. The nodes are added to the two routes in alternating order, first to the one route and then to the other. After allocating all nodes, the second route is reversed and connected onto the end of the first half of the route to ensure that the depot is visited first and last. Again, all nodes are visited by a truck.
- The third solution is generated using two parallel NNHs, as in the second solution. However, in the third solution, nodes are continuously added to either route depending on which endpoint is nearest to any unallocated node. The second route is again reversed after all nodes are allocated and connected to the first route to construct one full route where the depot is visited at the start and end of the route with truck-only nodes.
- The fourth solution is generated using an *alternating nearest neighbour heuristic* (ANNH). The first node, which is the depot, is allocated to the truck. Thereafter, the node with the smallest distance to the first node is allocated to a drone. The nearest node to the drone node is then chosen and allocated to the next free drone. This procedure is continued until all the drones have been assigned, after which a node is allocated to the truck again and the drone

assignments restart. This process repeats until all the nodes are assigned and the last node is again allocated as the depot.
- The fifth solution is generated using two parallel ANNHs, combining solutions two and four, by allocating nodes to the truck and thereafter, to all the drones before again allocating nodes to the truck. This process is continued until all nodes have been added. The second half of the route is again reversed and added onto the end of the first half of the route to ensure that the depot is visited first and last.
- The sixth solution is generated similarly to the third solution, except that nodes are now allocated to the truck and thereafter, to all the drones before again allocating a node to the truck (thus combining solutions three and four).

4 Problem Description

The TSPmDi considers a delivery system in which a single truck carries multiple drones to serve a set of customer locations. Drones can rejoin the truck either at a customer node or by intercepting the truck while the truck is en-route [10]. An example network and feasible route for the TSPmDi with two drones is shown in Fig. 1.

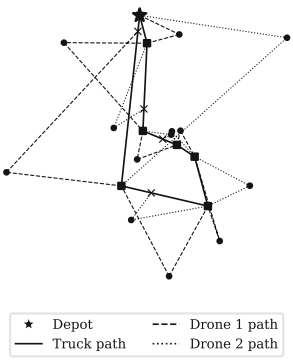


Fig. 1. TSPmDi example.

4.1 Network

The problem is defined on a complete, undirected graph, where the nodes represent the depot and customer locations, and edges represent feasible travel paths for both the truck and the drones. Distances between customer nodes are calculated using Euclidean distance. Drones are assumed to fly in straight lines between locations. The truck operates without road network constraints.

4.2 Vehicles

The system consists of one truck carrying a fixed number of identical drones. Each drone can carry a single package and is assumed to have unlimited range. After completing a delivery, a drone must return to the truck. Both the truck and drones travel at constant speeds, with drones assumed to move at twice the speed of the truck.

4.3 Customers

Each customer must be served exactly once, either by the truck or by a drone. A truck node is defined as a customer node that is assigned to be serviced by the truck. A drone node is defined as a customer node that is assigned to be serviced by a drone.

4.4 Routes

The delivery path consists of a truck route and a set of drone sub-tours. The truck route departs from the depot, visits all the assigned truck nodes, and returns to the depot. Drone sub-tours occur when a drone is launched from the truck to service an assigned drone node before returning to the truck. The vehicle routes, including the truck route and all drone sub-tours, and the assignment of customer nodes to vehicles are determined as part of the optimisation process. Figure 2 illustrates a drone sub-tour with an interception.

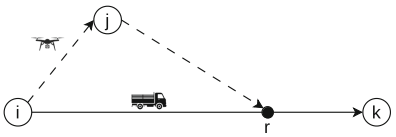


Fig. 2. Drone sub-tour with interception [10].

Multiple drones can be launched from the same truck node, and the truck may retrieve multiple drones either at a node or while travelling between nodes. After servicing a customer, a drone may either return directly to the next truck node or intercept the truck along its route.

Interception occurs when a drone's travel time from node i to j to k, denoted T_{ijk}, is less than the truck's travel time from node i to k, denoted T_{ik}. In this case, the interception location is calculated such that the drone rejoins the truck along the truck route at a point, r, on the edge (i, k). If interception is not feasible, the drone flies directly to node k. The total drone sub-tour time for servicing node j while the truck travels from node i to k, T^D_{ijk}, is given by

$$T^D_{ijk} = T_{ij} + s + \min\left(T_{jk},\ t^{\text{int}}_{ijk}\right), \tag{1}$$

where s denotes the service time at a customer node, T_{jk} denotes the time for the drone to travel directly from node j to the truck node k, and t_{ijk}^{int} denotes the feasible interception time.

Waiting time occurs if the truck arrives at the customer node k before all the drones assigned to the edge (i, k) have completed their sub-tours. In this case, the truck is required to wait for the remaining drone(s) to arrive at node k. The waiting time per edge (i, k), w_{ik}, is defined as

$$w_{ik} = \max\left(0,\ \max_{j \in \mathcal{D}_{ik}} \left(T_{ijk}^{D} - (T_{ik} + s)\right)\right), \tag{2}$$

where $\mathcal{D}_{ik}$ is the set of customer nodes served by drones during truck edge (i, k).

4.5 Objective

The objective is to minimise the total completion time of the delivery route. This total time includes the truck's travel time, the service times at the customer nodes, and the waiting time incurred. The overall objective is defined as

$$\min Z = \sum_{(i,k) \in \mathcal{E}_T} T_{ik} + \sum_{k \in \mathcal{N}_T \setminus \{0\}} s + \sum_{(i,k) \in \mathcal{E}_T} w_{ik}, \tag{3}$$

where $\mathcal{E}_T$ is the set of edges (i, k) traversed by the truck route, $\mathcal{N}_T$ is the set of nodes visited by the truck, and w_{ik} is the maximum delay introduced by any drone operating on edge (i, k), as defined in Eq. 2.

5 Proposed Algorithm

The proposed algorithm integrates the NNHis initialisation strategy, a TSPmDi problem representation, and an interchangeable metaheuristic framework. An activity diagram of the algorithm is shown in Fig. 3.

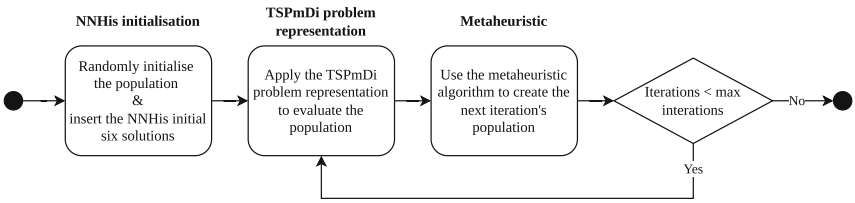


Fig. 3. Activity diagram of the proposed algorithm.

The algorithm is initialised by generating N real-valued vectors, each representing a candidate solution with $2(n)$ decision variables bounded between l_b and u_b, where n is the number of nodes excluding the depot. These decision variables

are iteratively updated throughout the optimisation process. To enhance initial solution quality, the NNHis strategy introduces six high-quality initial solutions to the initial population, constructed using nearest-neighbour heuristics based on truck-only routes and alternating truck-drone assignments.

Fitness evaluation is performed by mapping the continuous values to routes using priority-based mapping [38]. The first n values of each vector are sorted in descending order to determine the visiting sequence of the nodes. The remaining n values determine which vehicle services each node by applying either Solution Encoding 1 or Solution Encoding 2. Repair mechanisms are applied to ensure feasible solutions. The specific mapping and repair procedures for Solution Encoding 1 and Solution Encoding 2 are outlined in Sect. 5.1 and 5.2. This paper evaluates two metaheuristic algorithms: SaNSDE NNHis and GCPSO NNHis.

5.1 Solution Encoding 1

In Solution Encoding 1, the interval $u_b - l_b$ is divided into $d+1$ equal segments, where d is the number of drones in the system. Decision variable values in the first segment assign nodes to the truck, the second to drone 1, the third to drone 2, and so on. This encoding explicitly encodes the assignment of each node to a specific vehicle.

An infeasible solution occurs when a drone is assigned to more than one node between two consecutive truck nodes, as each drone can only serve one node in this interval. In such cases, the node remains a drone node but is reassigned to another available drone. If there are no available drones, then the node is reassigned to the truck. Importantly, repair operations involve simple reassignment and do not require re-sorting or re-discretisation of decision variables.

5.2 Solution Encoding 2

In Solution Encoding 2, the interval $u_b - l_b$ is divided into two equal segments. Values in the first segment assign nodes to the truck, while values in the second segment assign nodes to a drone. Unlike Solution Encoding 1, this encoding does not specify a specific drone at the assignment stage. A customer node that has been assigned to a drone may be assigned to any available drone.

An infeasible solution occurs when the number of consecutive drone-assigned nodes between two truck nodes exceeds the number of drones in the system, as each drone can serve only one node in this interval. In such cases, the $d+1^{th}$ node is reassigned to the truck to restore feasibility. This encoding separates node assignment from drone identity, significantly simplifying the mapping process and reducing encoding-induced infeasibility.

6 Empirical Evaluation

SE1 and SE2 are evaluated using 15 datasets from Bouman and Schmidt [39], featuring varying sizes and coordinate spreads. Single-centred datasets consist of

a central node around which most nodes are scattered; double-centred datasets contain two central locations with nodes distributed around both locations; and uniform datasets contain nodes that are spread out uniformly with no centre. Each distribution has five datasets with the following sizes: 10, 20, 50, 100, and 250 nodes.

Results for Solution Encoding 1 are taken directly from [10], while Solution Encoding 2 is introduced and evaluated in this paper. Both algorithms were run on each dataset for 30 independent runs of 1000 iterations each. The datasets are labelled as follows: '*distribution-reference-size*'. The uniform distributions are referred to as U, the single-centred distributions are referred to as SC, and the double-centred distributions are referred to as DC. For example, the label '*U-51-n10*' refers to the dataset with a uniform distribution, a reference number of 51, and a problem size of 10. The experiment parameters are provided in Table 2.

Table 2. Experiment parameters

Parameter	Value
Truck travel speed	10 distance units / time unit
Drone travel speed	20 distance units / time unit
Service time	0.1 time units
Population size (N)	100
Decision variable lower bound (l_b)	0
Decision variable upper bound (u_b)	200
Number of drones (d)	$2, 3, 4$

For Solution Encoding 2, the hyperparameters for the SaNSDE NNHis and GCPSO NNHis algorithms were tuned using Optuna, an automatic hyperparameter optimisation software framework [40]. The objective function was defined as the mean performance across all 15 datasets. A budget of 30 trials was allocated, and the *tree-structured parzen estimator* (TPE) sampler was employed to minimise the average cost function. The tuning was performed on the 2-drone configuration to establish a baseline set of robust parameters which were further evaluated on the 3 and 4-drone configurations to test the scalability of the algorithms. The hyperparameters of the SaNSDE NNHis and GCPSO NNHis algorithms are provided in Table 3 and Table 4.

6.1 SE1 vs. SE2 for the 2-Drone Configuration

The SaNSDE NNHis and the GCPSO NNHis algorithms were applied to both SE1 and SE2 respectively, for the 2-drone configuration. The means of the resulting delivery times and standard deviation across the 30 runs are presented in

Table 3. Hyperparameter search space and optimized values for SaNSDE NNHis

Parameter	Description	Search Range	Final Value
F	Differential weight	[0.1, 1.0]	0.991
C	Crossover probability	[0.1, 1.0]	0.801
p_{wr}	Mutation strategy probability	[0.1, 0.9]	0.162

Table 4. Hyperparameter search space and optimized values for GCPSO NNHis

Parameter	Description	Search Range	Final Value
c_1	Cognitive acceleration coefficient	[0.5, 3.0]	0.558
c_2	Social acceleration coefficient	[0.5, 3.0]	0.683
$v_{\max}$	Velocity clamping	[1.0, 10.0]	1.023

Table 5. The fastest delivery time per dataset is bolded. To assess the performance of the encodings across the datasets, a Wilcoxon signed-rank test was conducted. The test assessed whether there is a statistically significant difference between the aggregated results for SE1 and SE2 across all 15 datasets. As shown in Table 6, the test yielded a p-value of 0.5995, which is greater than the significance level of 0.05. Therefore, the results indicate that there is not a statistically significant difference between SE1 and SE2 for the two-drone configuration. With only two drones, the assignment space is relatively constrained, limiting the impact of encoding flexibility on solution quality. Figure 4 presents the final SE2 routes generated by the SaNSDE NNHis algorithm for the datasets with n = 20 to validate the legitimacy of the routes.

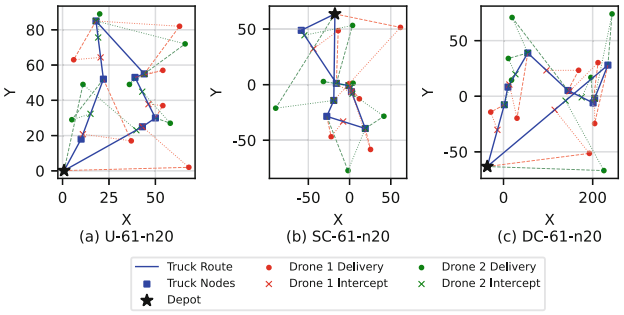


Fig. 4. SaNSDE NNHis routes for size 20 datasets.

6.2 Computational Performance: SE1 vs. SE2

This section analyses the computational performance of the two solution encodings, SE1 and SE2, across different spatial distributions and drone fleet sizes.

Table 5. Mean travel time and standard deviation for SE1 and SE2 for $d = 2$

Dataset	SE1 (Benchmarks)		SE2 (Results)	
	SaNSDE	GCPSO	SaNSDE	GCPSO
Uniform				
U-51-n10	**22.63 (0.15)**	23.32 (0.66)	22.66 (0.15)	22.86 (0.25)
U-61-n20	22.01 (0.36)	24.15 (0.48)	**21.79 (0.27)**	23.91 (1.15)
U-71-n50	52.40 (0.00)	**52.28 (0.36)**	55.72 (0.60)	53.18 (1.07)
U-91-n100	74.19 (0.00)	74.07 (0.29)	75.24 (0.85)	**73.77 (1.24)**
U-1-n250	120.12 (0.00)	**115.84 (0.24)**	119.45 (0.88)	117.09 (0.95)
Single centre				
SC-51-n10	23.10 (0.23)	24.18 (0.31)	**23.06 (0.10)**	24.48 (0.17)
SC-61-n20	30.82 (1.20)	31.74 (1.46)	**30.73 (0.80)**	35.47 (1.05)
SC-71-n50	65.24 (1.28)	58.39 (3.10)	52.93 (0.00)	**52.85 (0.46)**
SC-91-n100	123.71 (0.00)	119.43 (3.65)	**109.21 (3.17)**	111.38 (1.86)
SC-1-n250	192.73 (0.00)	185.14 (0.51)	181.28 (2.38)	**178.24 (3.65)**
Double centre				
DC-51-n10	33.63 (0.14)	34.15 (0.24)	**33.58 (0.00)**	34.67 (0.45)
DC-61-n20	59.07 (0.55)	60.46 (0.99)	**58.83 (0.93)**	62.26 (1.59)
DC-71-n50	114.43 (0.00)	111.22 (3.15)	**100.72 (0.00)**	102.49 (4.18)
DC-91-n100	153.38 (0.00)	151.29 (2.21)	140.43 (0.03)	**137.46 (2.69)**
DC-1-n250	249.76 (0.00)	**246.63 (4.00)**	258.64 (4.70)	249.97 (4.46)

The datasets with $n = 50$ are used for the computational evaluation. Table 7 reports the average CPU runtimes while Fig. 5 illustrates how runtime scales with the number of drones. Across all datasets, fleet sizes, and both metaheuristics applied, SE2 consistently outperforms SE1, with runtime reductions ranging from approximately 8% to over 17%. This performance gap is observed for both SaNSDE NNHis and GCPSO NNHis, indicating that it arises from structural differences between the encodings rather than algorithm-specific effects.

Empirical profiling results show that a dominant contributor to SE1's runtime is the discretisation step used to assign nodes to specific drones. SE1 maps decision values into $d + 1$ segments to assign specific drones, a process that is absent in SE2. For the $n = 50$ instances, this operation accounts for approximately $16-18\%$ of the total CPU runtime. This cost is incurred at every solution evaluation and remains constant across spatial distributions and fleet sizes.

SE2 eliminates this overhead entirely by replacing discretisation with a fixed interval assignment scheme, where the priority space is explicitly partitioned into one truck interval and one drone interval. As a result, SE2 avoids the repeated discretisation cost that dominates the computational profile of SE1.

To determine whether feasibility enforcement could offset this advantage, Fig. 6 reports the total number of repair operations executed by each encoding.

Table 6. Wilcoxon signed-rank test results for the aggregate performance comparison between SE1 and SE2 for $d = 2$

Metric	Value
Number of datasets (n)	15
Wilcoxon Statistic (W)	50.00
p-value	0.5995
Significance Level (α)	0.05
Result	**No Significant Difference**

Table 7. Average CPU runtime across 30 runs for $n = 50$ datasets

	2 Drones				3 Drones				4 Drones			
	SE1		SE2		SE1		SE2		SE1		SE2	
Instance	SaNSDE	GCPSO	SaNSDE	GCPSO	SaNSDE	GCPSO	SaNSDE	GCPSO	SaNSDE	GCPSO	SaNSDE	GCPSO
Uniform-71-n50	319.29	288.69	263.25	**246.59**	330.11	291.25	283.55	**273.23**	347.31	299.98	296.00	**248.48**
Singlecentre-71-n50	318.36	281.15	272.49	**255.81**	334.75	292.04	291.03	**270.53**	344.73	299.68	270.16	**256.08**
Doublecentre-71-n50	321.15	288.40	276.20	**260.11**	331.51	295.72	283.84	2 **65.57**	349.03	304.98	271.90	**261.80**

The results show that differences in feasibility repair behaviour do not explain the observed runtime differences.

For $d = 2$, SE2 occasionally produces more repairs than SE1, particularly under SaNSDE NNHis. The additional repairs occur because the coarse binary assignment (truck vs. drone), combined with random initialisation, increases the likelihood of exceeding drone capacity between consecutive truck nodes. However, these additional repairs are computationally inexpensive and do not outweigh the savings from removing discretisation.

As the fleet size increases to $d = 3$ and $d = 4$, a clear shift occurs. SE2 consistently generates fewer repairs than SE1 across all spatial distributions and both algorithms. This behaviour arises from the differing assignment mechanisms. For SE2, increasing the number of drones expands the feasible assignment space, providing greater flexibility and reducing forced reassignments to the truck. In contrast, SE1's discretisation-based mapping becomes increasingly sensitive as

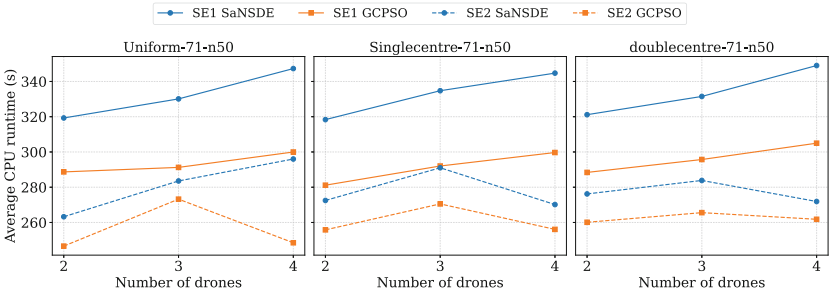


Fig. 5. Average CPU runtime across 30 runs for $n = 50$ datasets.

the number of drone intervals grows, leading to a higher probability of infeasible assignments and a corresponding increase in repair operations.

Overall, the results demonstrate that the primary driver of SE1's slower runtime is the fixed discretisation overhead, not the cost of feasibility enforcement. While repair behaviour varies with fleet size and distribution, repairs are either comparable between solution encodings or lower in SE2 for larger fleets. Consequently, SE2 remains consistently faster even in cases where SE2 performs more repairs, confirming that the elimination of discretisation results in the computational advantage.

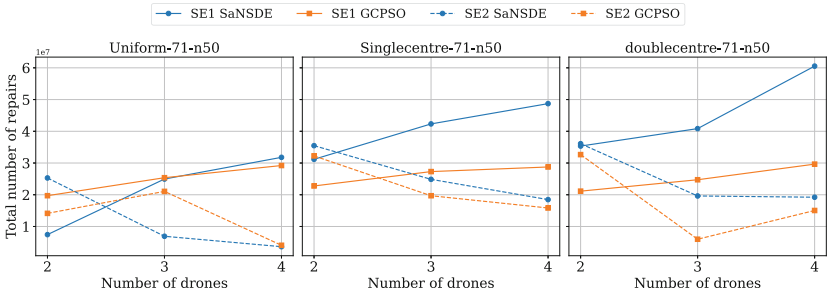


Fig. 6. Total repair mechanisms implemented across 30 runs for $n = 50$ datasets.

6.3 SE2 Scalability Evaluation

The SaNSDE NNHis and the GCPSO NNHis algorithms were applied to SE2 for fleet sizes $d = \{2, 3, 4\}$. The resulting means of the delivery times and standard deviation across the runs are presented in Table 8. An Iman-Davenport corrected Friedman test was conducted to evaluate performance across the six algorithm-fleet configurations. The results indicate a highly significant difference in performance across the six configurations ($F = 198.03, p < 0.001$), justifying the use of a post-hoc analysis.

A Nemenyi post-hoc test was performed to identify specific significant differences between algorithm-fleet pairs. The average ranks and significance groups for algorithm-fleet configurations ($d \in \{2, 3, 4\}$) are presented in Table 9. Configurations sharing a letter in the 'Group' column are not statistically different at $\alpha = 0.05$. While no significant difference was observed between SaNSDE NNHis and GCPSO NNHis at $d = 2$ ($p = 1.00$), both algorithms showed significant improvements as the fleet size increased ($p < 0.001$). Notably, SaNSDE NNHis with $d = 4$ achieved the highest overall ranking, though it was statistically indistinguishable from GCPSO NNHis with $d = 4$ and SaNSDE NNHis with $d = 3$ ($p > 0.05$).

For $d = 2$, the GCPSO NNHis is very competitive, winning in 7 out of 15 instances. For $d = \{3, 4\}$, SaNSDE NNHis has a better objective value for

Table 8. SE2 mean travel time and standard deviation for drone fleet sizes $d \in \{2, 3, 4\}$

Dataset	2 Drones		3 Drones		4 Drones	
	SaNSDE	GCPSO	SaNSDE	GCPSO	SaNSDE	GCPSO
Uniform						
U-51-n10	**22.66 (0.15)**	22.86 (0.25)	**19.36 (0.39)**	19.38 (0.35)	17.09 (0.00)	17.09 (0.00)
U-61-n20	**21.79 (0.27)**	23.91 (1.15)	**20.39 (0.56)**	21.36 (0.66)	**18.32 (0.29)**	18.65 (0.54)
U-71-n50	55.72 (0.60)	**53.18 (1.07)**	**36.29 (0.51)**	45.94 (1.27)	**33.19 (0.51)**	38.17 (0.30)
U-91-n100	75.24 (0.85)	**73.77 (1.24)**	**58.04 (0.94)**	63.94 (0.97)	**52.31 (1.05)**	59.87 (1.56)
U-1-n250	119.45 (0.88)	**117.09 (0.95)**	**103.79 (1.37)**	105.81 (2.09)	**93.36 (0.87)**	97.58 (0.9)
Single-centre						
SC-51-n10	**23.06 (0.10)**	24.48 (0.17)	21.05 (0.02)	21.05 (0.00)	**19.84 (0.02)**	19.87 (0.05)
SC-61-n20	**30.73 (0.80)**	35.47 (1.05)	**26.91 (0.31)**	28.38 (0.85)	**24.44 (0.71)**	25.49 (0.51)
SC-71-n50	52.93 (0.00)	**52.85 (0.46)**	**39.54 (0.73)**	44.35 (0.80)	**36.69 (0.70)**	43.93 (1.94)
SC-91-n100	**109.21 (3.17)**	111.38 (1.86)	**83.95 (2.09)**	95.79 (2.31)	**77.62 (4.16)**	89.45 (3.13)
SC-1-n250	181.28 (2.38)	**178.24 (3.65)**	**153.86 (3.03)**	161.51 (3.68)	**141.66 (1.88)**	153.60 (1.91)
Double-centre						
DC-51-n10	**33.58 (0.00)**	34.67 (0.45)	29.69 (0.00)	29.69 (0.00)	28.02 (0.00)	28.02 (0.00)
DC-61-n20	**58.83 (0.93)**	62.26 (1.59)	**54.60 (1.35)**	57.64 (1.73)	**48.51 (0.47)**	49.84 (1.22)
DC-71-n50	**100.72 (0.00)**	102.49 (4.18)	**84.69 (3.08)**	95.58 (1.84)	**76.18 (1.97)**	83.13 (3.47)
DC-91-n100	140.43 (0.03)	**137.46 (2.69)**	**108.39 (2.42)**	118.65 (3.30)	**95.92 (1.55)**	108.55 (2.97)
DC-1-n250	258.64 (4.70)	**249.97 (4.46)**	**208.58 (3.58)**	218.26 (2.65)	**193.92 (2.77)**	204.04 (6.49)

Table 9. Nemenyi post-hoc test results

Configuration	Average Rank	Group
SaNSDE NNHis_d4	1.33	A
GCPSO NNHis_d4	2.53	A, B
SaNSDE NNHis_d3	3.13	A, B
GCPSO NNHis_d3	4.00	C
SaNSDE NNHis_d2	5.00	D
GCPSO NNHis_d2	5.00	D

almost all datasets. However, the Nemenyi post-hoc test shows that there is no statistically significant difference between the SaNSDE NNHis and GCPSO NNHis algorithm per drone fleet size.

While GCPSO NNHis was initially identified as the best-performing algorithm for the $d = 2$ variant on large datasets in [10], SaNSDE NNHis is equally capable, and often numerically superior, when the drone fleet scales to $d = \{3, 4\}$.

7 Conclusion

This paper investigated the effect of solution encodings on the performance and the computational efficiency of continuous metaheuristic algorithms for the Travelling Salesman Problem with Multiple Drones and Interception. Two alternative

assignment strategies were evaluated within identical algorithmic frameworks, allowing the impact of encoding choice to be isolated from other algorithmic factors.

The experimental results show that SE2 consistently outperforms SE1 in terms of runtime across all tested datasets, drone fleet sizes, and optimisation algorithms, while producing solutions of comparable quality. This performance improvement is not explained by a lower number of repair operations. The dominant factor is the computational overhead introduced by the discretisation process required in SE1 to map continuous decision variables to specific drone indices. By avoiding this step, SE2 reduces per-iteration computational cost, even in cases where the number of repair mechanisms applied is similar or higher.

Increasing the drone fleet size consistently reduced the resulting travel time under SE2, although diminishing marginal returns were observed as the fleet size grew. Both metaheuristic algorithms produced statistically comparable results across the tested fleet sizes; however, SaNSDE NNHis yielded numerically superior solutions relative to GCPSO NNHis as the number of drones increased. Overall, these findings demonstrate that encoding-level design choices can have a decisive impact on algorithmic efficiency in continuous formulations of hybrid truck-and-drone routing problems.

Future research may investigate the incorporation of additional heuristic operators within the proposed algorithms. A road network can be included to investigate a real-world application.

Disclosure of Interests. The authors have no competing interests with the work discussed in this paper.

References

1. Oxford Dictionary of E-commerce. Oxford University Press (2020)
2. Chung, S.-H., Tsai, W.-H.: Drone-assisted last-mile delivery. Robot. Auton. Syst. **130**, 103–115 (2020)
3. Chang, C., Park, D.: Truck-drone collaborative delivery systems. In: Proceedings of the International Conference on Logistics, pp. 45–52 (2018)
4. 7-eleven just made the first commercial delivery by drone. https://www.theverge.com/2016/7/23/12262468/7-11-first-retailer-deliver-food-drone Accessed 11 May 2025
5. Mckinsey and Company. Drones take to the sky, potentially disrupting last-mile delivery. https://www.mckinsey.com/industries/aerospace-and-defense/our-insights/future-air-mobility-blog/drones-take-to-the-sky-potentially-disrupting-last-mile-delivery Accessed 11 May 2025
6. Mckinsey and Company. Consumer views of drone delivery: How soon can you get here? https://www.mckinsey.com/industries/aerospace-and-defense/our-insights/future-air-mobility-blog/consumer-views-of-drone-delivery-how-soon-can-you-get-here Accessed 11 May 2025
7. Moshref-Javadi, M., Winkenbach, M.: Applications and research avenues for drone-based models in logistics: a classification and review. Expert Syst. Appl. **177**, 114854 (2021)

8. Agatz, N., Bouman, P., Schmidt, M.: Optimization approaches for the traveling salesman problem with drone. Transp. Sci. **52**(4), 965–981 (2018)
9. Moremi, T.J., Grobler, J., Kaminsky, P.M.: An ant colony optimisation approach to scheduling truck and drone delivery systems, PhD thesis, Stellenbosch University (2022)
10. Dillon, S., Grobler, J.: Solving routing problems with multiple drones with interceptions. In: Fourth Decision Science Alliance International Summer Conference. Springer (2026)
11. Boccia, M., Mancuso, A., Masone, A., Murino, T., Sterle, C.: New features for customer classification in the flying sidekick traveling salesman problem. Expert Syst. Appl. **247**, 123106 (2024)
12. Chung, S.H., Sah, B., Lee, J.: Optimization for drone and drone-truck combined operations: a review of the state of the art and future directions. Comput. Operat. Res. **123**, 105004 (2020)
13. Liang, Y., Luo, Z.: A survey of truck-drone routing problem: literature review and research prospects. J. Operat. Res. Soc. China **10**(2), 343–377 (2022)
14. Madani, B., Ndiaye, M.: Hybrid truck-drone delivery systems: a systematic literature review. IEEE Access **10**, 92854–92878 (2022)
15. Otto, A., Agatz, N., Campbell, J., Golden, B., Pesch, E.: Optimization approaches for civil applications of unmanned aerial vehicles (uavs) or aerial drones: a survey. Networks **72**(4), 411–458 (2018)
16. Poikonen, S., Campbell, J.F.: Future directions in drone routing research. Networks **77**(1), 116–126 (2021)
17. Zhang, R., Dou, L., Xin, B., Chen, C., Deng, F., Chen, J.: A review on the truck and drone cooperative delivery problem. Unmanned Syst. **12**(05), 823–847 (2024)
18. Boysen, N., Briskorn, D., Fedtke, S., Schwerdfeger, S.: Drone delivery from trucks: drone scheduling for given truck routes. Networks **72**(4), 506–527 (2018)
19. Ha, Q.M., Deville, Y., Pham, Q.D., Hà, M.H.: On the min-cost traveling salesman problem with drone. Trans. Res. Part C: Emerging Technol. **86**, 597–621 (2018)
20. Knoetze, F.E.: Solving a last mile truck and drone delivery schedule to optimality. University of Stellenbosch, Final year project (2021)
21. Ernst, R.: Evolutionary algorithms for routing problems with drones with interceptions, PhD thesis, Stellenbosch University (2024)
22. Betti Sorbelli, F., Corãš, F., Das, L.P., Pinotti, C.M.: Greedy algorithms for scheduling package delivery with multiple drones. In: ACM International Conference Proceeding Series, pp. 31–39 (2022)
23. Sorbelli, F.B., Corãš, F., Das, S.K., Palazzetti, L., Pinotti, C.M.: On the scheduling of conflictual deliveries in a last-mile delivery scenario with truck-carried drones. Pervasive Mob. Comput. **87** (2022)
24. Salama, M.R., Srinivas, S.: Collaborative truck multi-drone routing and scheduling problem: package delivery with flexible launch and recovery sites. Trans. Res. Part E: Logist. Trans. Rev. **164** (2022)
25. Han, B., Chi, T., Sun, Z.: Collaborative delivery problem of trucks and multiple drones: an en-route operation with flexible launch positions. Int. Trans. Oper. Res. (2025)
26. Lin, M., Chen, Y., Han, R., Chen, Y.: Discrete optimization on truck-drone collaborative transportation system for delivering medical resources. Discret. Dyn. Nat. Soc. (2022)
27. Lin, M., Lyu, J.-Y., Gao, J.-J., Li, L.-Y.: Model and hybrid algorithm of collaborative distribution system with multiple drones and a truck. Sci. Program. (2020)

28. Narayan, S., Paulos, J., Chen, S.W., Manjanna, S., Kumar, V.: Combined routing and scheduling of heterogeneous transport and service agents. In: IEEE International Conference on Intelligent Robots and Systems, pp. 1466–1472 (2021)
29. Li, H., Chen, J., Wang, F., Zhao, Y.: Truck and drone routing problem with synchronization on arcs. Nav. Res. Logist. **69**(6), 884–901 (2022)
30. Thomas, T., Srinivas, S., Rajendran, C.: Collaborative truck multi-drone delivery system considering drone scheduling and en route operations. Ann. Oper. Res. **339**(1–2), 693–739 (2024)
31. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential evolution: a practical approach to global optimization. Springer (2005)
32. Yang, Z., Tang, K., Yao, X.: Self-adaptive differential evolution with neighborhood search. In: 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), pp. 1110–1116. IEEE (2008)
33. Qin, A.K., Suganthan, P.N.: Self-adaptive differential evolution algorithm for numerical optimization. In: 2005 IEEE congress on evolutionary computation, vol. 2, pp. 1785–1791. IEEE (2005)
34. Kennedy, J., Eberhart, R.:Particle swarm optimization. In: Proceedings of ICNN 1995-International Conference on Neural Networks, vol 4, pp. 1942–1948. IEEE (1995)
35. Sahab, M.G., Toropov, V.V., Gandomi. A.H.: A review on traditional and modern structural optimization: problems and techniques. Metaheuristic Appli. Struct. Infrastruct., 25–47 (2013)
36. van den Bergh, F., Engelbrecht, A.P.: A new locally convergent particle swarm optimiser. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 3, p. 6 IEEE (2002)
37. Dillon, S., Ernst, R., Grobler, J.: Investigating alternative clustering algorithms for a cluster-first, route-second truck and drone scheduling algorithm. In: Gervasi, O., et al. ICCSA 2023. LNCS, vol. 13957. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36808-0_10
38. Grobler, J., Engelbrecht, A.P., Kok, S., Yadavalli, S.: Metaheuristics for the multi-objective fjsp with sequence-dependent set-up times, auxiliary resources and machine down time. Ann. Oper. Res. **180**(1), 165–196 (2010)
39. Bouman, P., Agatz, N., Schmidt, M.: Instances for the tsp with drone (2015). https://zenodo.org/records/22245
40. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)

Shared Autonomy for an Omnidirectional Robot with SLAM and Collision Avoidance

Mireya Zapata[1(✉)], Alejandro Camino[1], Vannesa Vargas[2], and Pablo Ramos[2]

[1] Centro de Investigación en Mecatrónica y Sistemas Interactivos (MIST), Facultad de Ingenierías, Universidad Indoamérica, Machala y Sabanilla, Quito, Ecuador
mireyazapata@uti.edu.ec

[2] Departamento de Eléctrica, Electrónica y Telecomunicaciones, Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui y Ambato, 171103 Sangolquí, Ecuador

Abstract. This paper describes the design and testing of an omnidirectional mobile robot (OMR) with Mecanum wheels, designed to implement assisted teleoperation within the Robot Operating System (ROS) framework. The platform has a 2D LiDAR sensor for environmental perception. Which enables real-time mapping and localization; in this case using the Hector SLAM algorithm to build 2D occupancy maps in real time, along with the Dynamic Window Approach (DWA) to ensure a safety layer that avoids obstacles while remote-controlling. The proposed shared-autonomy architecture demonstrates that combining SLAM-based mapping with reactive navigation makes teleoperation tasks much safer and more efficient. Experimental results show that the system allows the operator to control the mobile robot while still enabling it to avoid collisions on its own. The system is designed to support, not override, the operator's commands; instead, it only takes action when in close proximity with an obstacle. In contrast with other obstacle avoidance solutions, our proposal allows the user to retain navigation control as the obstacle avoidance algorithm takes action, additionally it supports real-time mapping; this allows for navigation on unknown scenarios while maintaining obstacle avoidance capabilities. Validation of the prototype proves the efficacy of this hybrid methodology for assisted teleoperation applications. The prototype shows high odometry precision (relative error $<1\%$) and high mapping accuracy, demonstrating the system's robustness. Furthermore, testing shows that autonomous safety features in a teleoperation system, illustrating this hybrid methodology as a feasible solution for applications requiring accuracy and safety in intricate environments.

Keywords: Omnidirectional robot · mecanum wheel · collision avoidance · ROS · SLAM · Hector SLAM · teleoperation · shared autonomy · Dynamic Window Approach

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 556–567, 2027.
https://doi.org/10.1007/978-3-032-30494-0_35

1 Introduction

Mobile robotics has witnessed significant growth, with an increasing variety of autonomous robots capable of carrying out complex tasks in diverse environments such as home, industry, military, and rescue scenarios. The ability of these systems to sense, learn, and engage with the world around them is a key enabling technology for recent advances in control, and robotic systems. Nevertheless, path planning in an unknown, dynamic environment remains a challenging problem for the increased autonomy and intelligence of mobile robots, due to a number of constraints imposed by the robot and its environment. The uncertainty characterizing real-world environments and the inherent limitations of sensor technology make the design of robust and intelligent navigation systems a major challenge today. In addition, the generation of environment maps is necessary for a wide range of applications, such as vehicle navigation, robot obstacle avoidance, search and rescue operations, and indoor and outdoor surveillance [1,2].

In this sense, omnidirectional mobile robots (OMRs) are a novel class of robots that can freely move in any direction and at any orientation, defying the non-holonomic constraints of traditional robots which need to reorient to change direction. This feature provides better maneuverability and accuracy, which is very useful in a limited-space working area and in a cluttered or dynamic environment, e.g., factory workshops, warehouses, or hospitals [3].

Teleoperation allows the human operator to directly control a robot from a remote location, being very effective in situations where the environment is hazardous or unreachable for human beings. In these architectures, the human performs high-level cognitive and localization tasks, while the robot handles low-level motion control. Shared autonomy methodologies for teleoperation have recently overcome key challenges such as network delays and model errors. Gulecyuz et al. [4] introduced a new shared autonomy mechanism that learns from teleoperated demonstrations during late-stage deployment, despite network-induced delays, and using confidence-aware arbitration balancing human and autonomous control. Likewise, Jabbour [5] proposed a two-term adaptive assistance scheme for compensating for model uncertainties, where operator input is considered as a corrective measurement and by integrating adaptive Kalman filtering in real-time for compensation of local errors and Procrustes analysis for correcting global misalignment.

Behind the autonomy and transparency of teleoperation, there is the importance of the real-time pose tracking and map-building system capabilities. The simultaneous localization and mapping (SLAM) problem is fundamental in mobile robotics and enables a robot to construct a representation (map) of an unknown environment while concurrently tracking its location and orientation within that representation [2]. LiDAR sensors have become increasingly popular for SLAM, mainly due to their accuracy, robustness, wide field of view, and insensitivity to light conditions. Recent research has analyzed different 2D LiDAR SLAM algorithms, exemplifying the trade-off between filter-based methods, e.g.,

Hector SLAM and Gmapping, and graph-based methods such as Cartographer and KartoSLAM [6,7].

Moreover, reactive obstacle avoidance is essential for safe navigation in dynamic and rough environments, particularly when robots come in contact with humans. Reactive control strategies enable the robot to adapt its path while traveling based on real-time sensor information to avoid collisions [8]. In this sense, there has been a rising interest on local planning algorithms such as the Dynamic Window Approach (DWA) taking into account its computational efficiency and kinodynamic feasibility [9]. Recent extensions of DWA have been introduced to address its drawbacks in more complex scenarios. Zhang et al. [10] proposed an enhanced DWA based on SSA for UAV path planning in a multi-dynamic obstacle environment, where the path length, safety distance, and iteration efficiency were notably improved. However, integrating real-time LiDAR-SLAM perception with the kinematic control of omnidirectional mobile robots (OMR) in dynamic environments remains a worthwhile problem to be investigated [11]. Sensor fusion solutions have proven effective for overcoming localization issues in indoor environments without GPS signals. Lee et al. [12] showed that wheel-visual-inertial odometry outperforms other standalone odometry methods in feature-poor environments.

In this paper, the challenge of providing a dependable and secure robotic system in a cluttered and dynamic environment is addressed by developing an OMR control software using ROS. The purpose is to build a teleoperation platform that couples real-time localization and mapping system with reactive obstacle avoidance for safe navigation under human control, reconciling the autonomy efficiency with the intelligent supervision of a human operator.

The remainder of this paper is organized as follows. Section 2 details the hardware and software architecture of the implemented system, along with an explanation of the proposed experiments to validate the system's performance. Section 3 presents the results obtained in the evaluation process, presenting real-life testing data and providing visual representations of the experiments. Section 4 discusses the results and its limitations. Finally, Sect. 5, concludes the work and analyzes how the proposed solution contributes to teleoperation-obstacle avoidance sector.

2 Methods

The following section presents the applied research methodology for this project, divided into four major parts: hardware platform implementation, software architecture design, system integration, and experimental validation.

2.1 Hardware Platform Implementation

The mobile robot, shown in Fig. 1, is an OMR with four Mecanum wheels (MD60, 152 mm diameter), each powered by a 24 V DC motor (MD60P18P, 100 W, 175 rpm nominal) with a quadrature encoder.

The four RoboClaw Solo 30A motor controllers manage the low-level PID velocity control for the four motors, with gains adjusted using the BasicMicro Motion Studio software.

A Raspberry Pi 4 Model B+ (4 GB RAM) serves as the main computer on board, running Ubuntu Server 20.04 LTS and ROS Noetic. Environmental perception is provided by an RPLidar A2M8 2D LiDAR sensor (16m range, 8000 samples/s).

The system is powered by a 60 V, 20 Ah lithium main battery, stepped down to 24 V for the motors and 5 V for the logic components. A wireless joystick allows remote operation.

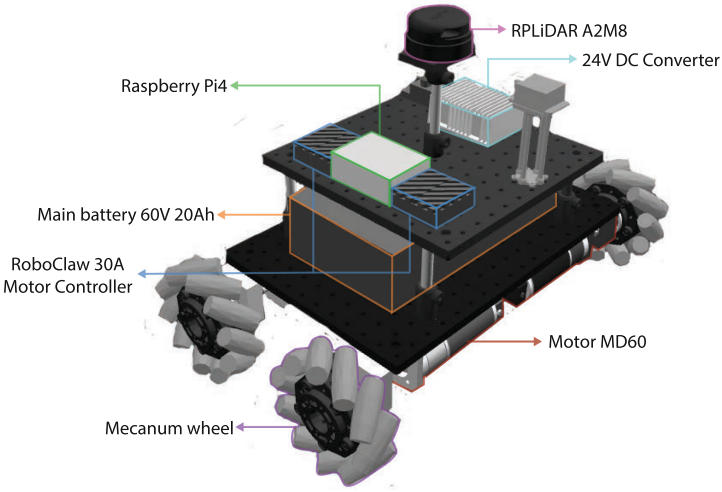


Fig. 1. Prototype of the omnidirectional mobile robot and its 3D model. Physical prototype with motor labels (M1-M4).

2.2 Software Architecture and ROS Implementation

The software architecture, shown in Fig. 2, is a modular ROS-based system. The key nodes and their interactions are as follows:

- **rplidarNode.cpp**: Interfaces with the RPLidar sensor, publishing raw laser scan data to the `/scan` topic (sensor_msgs/LaserScan).
- **joystick_bypass_dwa.py**: Captures joystick commands and publishes them as raw velocity commands on the **`/cmd_vel_joystick`** topic (geometry_msgs/Twist).
- **advanced_safety_mini_dwa.py**: Implements the shared autonomy core. It subscribes to `/scan` and **`/cmd_vel_joystick`**, using the DWA algorithm

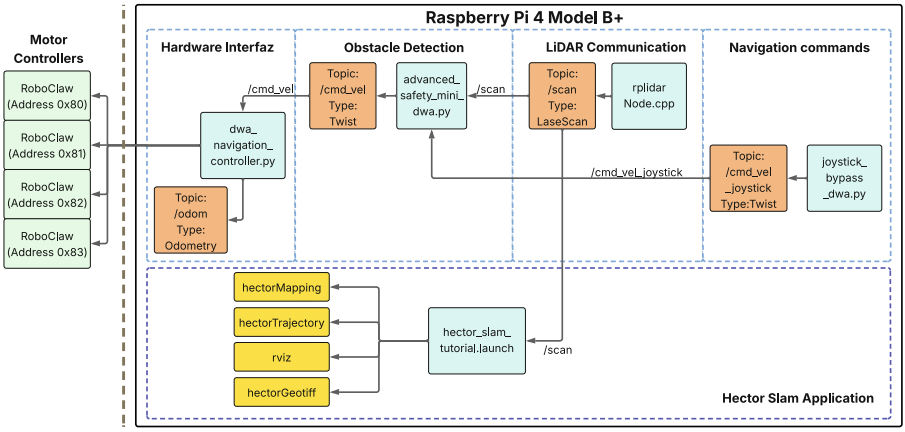


Fig. 2. ROS software architecture for the omnidirectional mobile robot, showing nodes (blue), topics (orange), and data flow. (Color figure online)

to generate safe velocity commands when in proximity with obstacles. These are published on the `/cmd_vel` topic, effectively filtering the operator's commands to avoid imminent collisions.

- **dwa_navigation_controller.py**: Subscribes to `/cmd_vel`, computes the inverse kinematics for the four Mecanum wheels based on Eq. 1, and sends individual velocity setpoints to each RoboClaw controller via serial communication. It also computes odometry from wheel encoders and publishes it on the `/odom` topic (nav_msgs/Odometry).
- **hector_mapping** package: Subscribes to the `/scan` topic to perform SLAM, generating a 2D occupancy grid map and publishing the robot's estimated pose (`/slam_out_pose`). This map is visualized in RViz.

The inverse kinematics that relates the robot's velocity vector $\mathbf{U} = [v_x, v_y, \omega]^T$ to individual wheel angular velocities ω_i is given by:

$$\begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \end{bmatrix} = \frac{1}{r} \begin{bmatrix} 1 & -1 & -(l_x + l_y) \\ 1 & 1 & l_x + l_y \\ 1 & -1 & l_x + l_y \\ 1 & 1 & -(l_x + l_y) \end{bmatrix} \begin{bmatrix} v_x \\ v_y \\ \omega \end{bmatrix} \tag{1}$$

where r is the wheel radius, and l_x and l_y are the distances from the robot's center to the wheel axles along the x and y axes.

An important clarification of the proposed system needs to be established. As mentioned previously the obstacle avoidance layer is designed to support the operator and not completely override the operator's command. In this context, the **advanced_safety_mini_dwa.py** node receives the raw velocity command of the joystick `/cmd_vel_joystick`, and as long as this node does not detect an obstacle inside the threshold region, the raw joystick command is directly passed to the final output velocity command `/cmd_vel`. In the case that the algorithm

detects an obstacle inside the region, the joystick command is processed to represented a desired pose instead of a velocity command, taking the direction of the desired joystick velocity and projecting a point in this direction; this pose is the target position that the DWA algorithm takes as reference to compute its velocity command.

2.3 Experimental Validation

Three sets of experiments were conducted to validate the system's performance:

1. **Odometry Accuracy**: The robot was programmed to perform linear displacements of 400 cm in the x and y directions (10 repetitions each) and rotational maneuvers ($90°, 180°, 270°, 360°$) with five repetitions each. Odometry estimates were compared against real displacements and IMU data (for rotations) to compute absolute and relative errors.
2. **Mapping Precision**: Three independent mapping runs of an L-shaped corridor (dimensions specified in Fig. 3) were performed using Hector SLAM. The lengths of eight distinct segments in the generated maps were measured using RViz's measure tool and compared to ground-truth manual measurements.
3. **Obstacle Avoidance Effectiveness**: The robot was teleoperated directly towards a static obstacle from varying initial distances (2.2 m to 4.2 m). Each repetition measured the time taken by the DWA algorithm to avoid the obstacle interpreted as the time difference between the algorithm taking control of the velocity command to it returning full control to the operator.

3 Results

The experimental results are presented below, focusing on the three key performance metrics.

3.1 Odometry Accuracy

The odometry system demonstrated high precision. For linear displacements, the average errors were minimal, as shown in Table 1. For rotations, all relative errors remained below 1%, with the highest error (0.62%) observed for the 360° rotation. The results confirm the robustness of the odometry estimation for navigation tasks.

3.2 Cartography Accuracy with Hector SLAM

To assess the accuracy of the maps, three independent mappings of an L-shaped corridor were conducted. The experimental space layout with eight rectilinear segments a-h (see Fig. 3). The lengths of the segments were measured using the RViz-measure tool and compared with manual ground-truth measurements. The full details for each segment across the three runs are given in Table 2, and Fig. 4 shows one of the resulting occupancy grid maps in RViz.

Table 1. Absolute and relative errors for linear and angular odometry.

Test Type	Real Value	Mean Value	Rel. Error (%)
Δx	400 cm	401.7 cm	0.42
Δy	400 cm	403.8 cm	0.95
$\Delta\theta$ (90°)	90°	90.6°	0.43
$\Delta\theta$ (180°)	180°	180.6°	0.11
$\Delta\theta$ (270°)	270°	270.3°	0.32
$\Delta\theta$ (360°)	360°	358.4°	0.62

These results show great metric accuracy in the cartographic reconstruction. Analyzing the eight measured segments, the relative error is always less than 0.35%, with a mean value of 0.19%. Although the largest absolute errors are attained for the two longest segments (d and h, with –1.70 cm and –2.20 cm, respectively), the corresponding relative errors are still small (0.17% and 0.13%), so the important conclusion is that the magnitude of the error is scaled with the measured distance.

The reproducibility of the three separate runs for each segment demonstrates the robustness of the mapping, confirming that Hector SLAM produces metrically accurate environment maps that can be used in an application with a mobile robot.

These results are consistent with Zhao et al. [7], who showed that the mapping accuracy of Hector SLAM in structured environments is very high.

Table 2. Measurements of the eight segments (a-h) within the mapped space, as obtained from RViz.

Segment of the path	Ground Truth (cm)	Map Measurements			Mean value (cm)	Rel. Error (%)
		1	2	3		
a	188.2	187.6	187.6	187.5	187.6	0.34
b	179.8	178.6	179.5	180.0	179.4	0.24
c	221.0	220.9	220.7	220.6	220.7	0.12
d	1011.3	1010.3	1007.6	1010.9	1009.6	0.17
e	178.4	177.7	178.4	178.3	178.1	0.15
f	176.5	176.4	175.6	176.3	176.1	0.23
g	176.5	176.2	177.1	176.8	176.7	0.11
h	1747.5	1745.9	1745.3	1744.7	1745.3	0.13

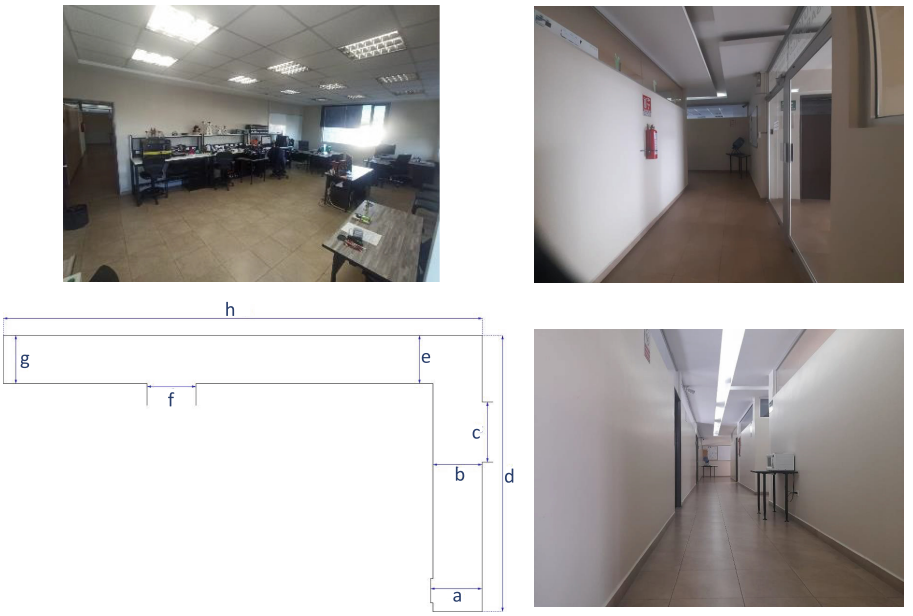


Fig. 3. Layout of the L-shaped corridor used as the testing scenario. The segments labeled with letters (from a to h) indicate the locations where measurements were taken during the tests.

3.3 Obstacle Avoidance with DWA

The DWA-based safety layer consistently intervened to prevent collisions. The reaction time of the algorithm was stable, averaging around 5.36 s across four trials with varying initial distances (see Table 3). This consistency suggests the avoidance behavior is governed by the fixed safety parameters (e.g., 0.8 m safe distance) rather than the initial distance. Figure 5 shows a teleoperation session where the DWA algorithm adjusted the commanded velocity in real time to avoid an obstacle, while Hector SLAM continued to build the map.

The observed performance is consistent with recent DWA enhancements reported in the literature. Zhang et al. [13] demonstrated that incorporating gradient information into DWA improves obstacle awareness, particularly for non-convex obstacles. While our implementation uses the standard DWA formulation, the consistent reaction times validate its suitability as a non-intrusive safety layer for shared autonomy applications.

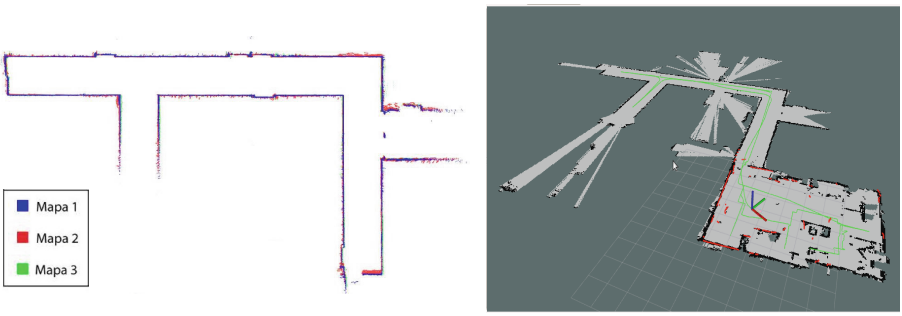


Fig. 4. Occupancy grid map of the L-shaped corridor generated by Hector SLAM, visualized in RViz. The overlay of the three obtained maps demonstrates the reproducibility of the measurements.

Table 3. Reaction time of the DWA algorithm for obstacle avoidance from varying initial distances.

Trial	Initial Distance (m)	DWA Reaction Time (s)
1	2.20	5.37
2	2.20	5.50
3	3.20	5.16
4	4.20	5.40

4 Discussion

The results confirm the viability of the shared-autonomy approach for the Mecanum-wheeled robot. The strongly coupled integration of Hector SLAM and a DWA-based safety layer in a ROS environment was successful on performing assisted teleoperation.

The high precision of the odometry system (relative errors <1%) is noteworthy, as it mitigates one of the primary drawbacks of Mecanum wheels—slippage. This accuracy is crucial for reliable localization within the SLAM framework. Similarly, the mapping accuracy (0.19% mean relative error) confirms that Hector SLAM, using only LiDAR data, can generate metrically consistent maps of indoor environments, providing valuable situational awareness to the remote operator. The consistent performance of the DWA algorithm, with its reaction times, validates its use as a non-intrusive safety layer. It effectively acts as an "assistant," modifying the operator's commands only when necessary to prevent a collision, thereby reducing cognitive load and enhancing safety without completely removing human control. This is clearly evidenced in Fig. 5, where the

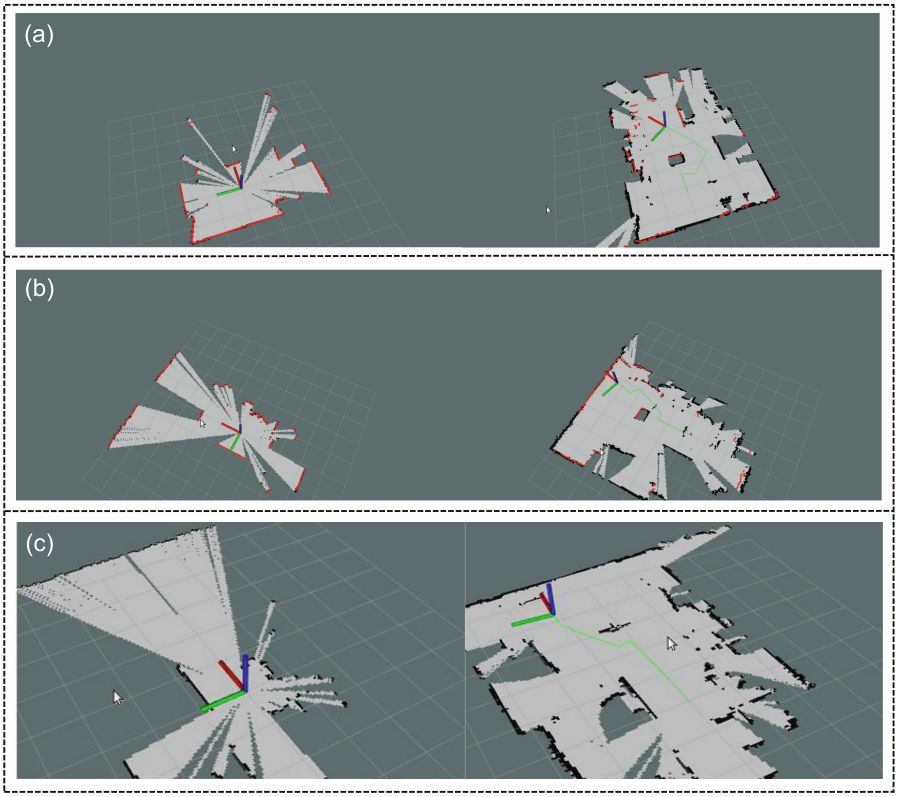


Fig. 5. Experimental validation of obstacle avoidance with DWA in RViz. (a) System initiates mapping under manual joystick control while ensuring safe navigation. (b-c) In obstacle-dense environments, DWA adjusts trajectories in real time, maintaining safe and fluid paths, while Hector SLAM generates coherent and detailed maps. (Color figure online)

executed trajectory (green line) smoothly deviates from the direct path to avoid an obstacle detected by the LiDAR (red points).

However, the study has limitations. The inherent susceptibility of Mecanum wheels to slippage on uneven terrains remains a challenge for odometry, though the use of SLAM partially corrects for this. Hector SLAM, while robust, can drift during rapid rotations, and the 2D LiDAR is blind to obstacles above or below its scan plane, such as table overhangs. Additionally for the DWA implementation, the velocity command transition logic using distance threshold might not be sufficient in the presence of dynamic obstacles. Moreover, although the conversion from joystick velocity command to reference pose seemed to be sufficient in the testing environment of this study, it might not prove so efficient in more diverse and cluttered environments, as the forward reference point is

set at a constant distance from the robot it could place the pose inside of an obstacle.

Future work will focus on sensor fusion by integrating an IMU and depth cameras for more robust localization and perception. There is room for improvement on the logic for when to activate the obstacle avoidance layer, future work fill focus on smoother and safer transition. Furthermore by improving the transformation from joystick command to forward reference point the performance of the DWA algorithms would highly improve. Besides that, migrating to ROS 2 would also provide better real-time performance and support for more complex, multi-sensor fusion algorithms and localization algorithms.

5 Conclusion

In this paper, we have presented the design, implementation, and validation of an omnidirectional mobile robot, based on Mecanum wheels, for shared autonomy teleoperation in the ROS environment. The main contribution is the tight integration of Hector SLAM for real-time mapping, with a minimally invasive reactive obstacle avoidance layer based on DWA, that plays the role of non-intrusive safety assistant for a human operator.

Experimental results verified that the system can simultaneously obtain high precision odometry and mapping (with relative errors less than 1% and 0.2%, individually), and perform efficient collision avoidance. The results demonstrate that this hybrid shared autonomy approach improves the safety and efficiency of teleoperation in challenging scenarios, providing a feasible solution to practical applications such as remote inspection in hazardous regions or high-precision logistics. The ROS implementation is open source, providing a solid basis for future research and development in autonomous mobile robots.

References

1. Li, Z., Chamran, K., Alobaedy, M.M., Sheikh, M.A., Siddiqui, T., Ahad, A.: An optimization on 2d-slam map construction algorithm based on lidar. J. Intell. Robotic Syst. **110**(4) (2024). https://doi.org/10.1007/s10846-024-02123-1
2. Payá, L., Gil, A., Reinoso, O.: A state-of-the-art review on mapping and localization of mobile robots using omnidirectional vision sensors. J. Sensors **2017**, 1–20 (2017). https://doi.org/10.1155/2017/3497650
3. Burkacki, M., Łysy, I., Suchoń, S., Chrzan, M., Kowolik, R.: Systematic review of mecanum and omni wheel technologies for motor impairments. Appl. Sci. **15**(9), 4773 (2025). https://doi.org/10.3390/app15094773
4. Güleçyüz, B.: Enhancing shared autonomy in teleoperation under network delay: transparency- and confidence-aware arbitration. IEEE Robot. Autom. Lett. **10**(10), 9654–9661 (2025). https://doi.org/10.1109/lra.2025.3596436
5. Elio, J.: Shared-autonomy control for improving Human-Robot collaboration in haptic teleoperation. Theses, Université de Bordeaux (Nov 2025). https://theses.hal.science/tel-05460108

6. Ran, Y., Xu, X., Tan, Z., Luo, M.: A review of 2d lidar slam research. Remote Sensing **17**(7), 1214 (2025). https://doi.org/10.3390/rs17071214
7. Zhao, J., Liu, S., Li, J.: Research and implementation of autonomous navigation for mobile robots based on slam algorithm under ros. Sensors **22**(11), 4172 (2022). https://doi.org/10.3390/s22114172
8. Azizi, M.R., Rastegarpanah, A., Stolkin, R.: Motion planning and control of an omnidirectional mobile robot in dynamic environments. Robotics **10**(1), 48 (2021). https://doi.org/10.3390/robotics10010048
9. Lin, Z., Taguchi, R.: Faster implementation of the dynamic window approach based on non-discrete path representation. Mathematics **11**(21), 4424 (2023). https://doi.org/10.3390/math11214424
10. Zhang, L., Li, Y., Yu, Y., Zhao, Y., Song, W., Retscher, G.: Improved dynamic window approach for uav local path planning in multi-dynamic obstacles environments using sparrow search algorithm. Trans. Emerging Telecommun. Technol. **36**(10), e70284 (2025). https://doi.org/10.1002/ett.70284, https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.70284
11. Rivai, M., Hutabarat, D., Nafis, Z.M.J.: 2d mapping using omni-directional mobile robot equipped with lidar. TELKOMNIKA (Telecommun. Comput. Electr. Control) **18**(3), 1467 (2020). https://doi.org/10.12928/telkomnika.v18i3.14872
12. Lee, C., Hur, S., Kim, D., Yang, Y., Choi, D.: Insufficient environmental information indoor localization of mecanum mobile platform using wheel-visual-inertial odometry. J. Mech. Sci. Technol. **38**(9), 5007–5015 (2024). https://doi.org/10.1007/s12206-024-0836-z
13. Zhang, Z., Xue, Y., Figueroa, N., Åkesson, K.: Gradient field-based dynamic window approach for collision avoidance in complex environments. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 19669–19674 (2025). https://doi.org/10.1109/IROS60139.2025.11246091

Bi-objective Optimization of Electric Vehicle Charging Scheduling with Heterogeneous Chargers Under Time-Varying Electricity Prices

Imene Zaidi[1(✉)], Latifa Belhocine[1,2], Baptiste Buchi[1], and Hajar Nouinou[1]

[1] CESI LINEACT, UR 7527, Nancy, France
{izaidi,bbuchi,hnouinou}@cesi.fr, latifa.belhocine@mines-albi.fr
[2] Centre de Génie Industriel, IMT Mines Albi, Albi, France

Abstract. This paper addresses the optimization of charging schedules in EV charging stations with heterogeneous chargers under time-varying electricity prices. The objective is to decide which demands to accept and how to schedule them to maximize the number of demands served while minimizing electricity costs. We formulate the problem as a bi-objective model and compare two approaches: an exact ε-constraint method and a hybrid NSGA-II enhanced with repair mechanisms and refinement procedures. Computational results on 70 benchmark instances with up to 400 EVs and four price scenarios show that the hybrid NSGA-II produces well-distributed Pareto fronts and achieves competitive performance on small instances, while significantly outperforming the exact method on large instances, with improvements of up to +59.89% in normalized hypervolume and an average gain of +40.8% for instances with 300–400 EVs. In addition, it reduces computation time by up to 77%, with an average reduction of 65% for instances ranging from 50 to 400 EVs.

Keywords: Electric Vehicle · Charging Scheduling · Multi-Objective Optimization · Electricity Cost Optimization

1 Introduction

Although electric vehicles (EVs) emerged alongside the first automobiles, their adoption declined during the twentieth century before reemerging in the 2010s as part of the clean energy transition. According to the International Energy Agency [7], global EV sales reached 17 million units in 2024, representing over 20% of total car sales. Progress in EV battery technology is shortening charging times and could make them comparable to refuelling, provided that charging infrastructure expands quickly to support the growing adoption of EVs. In 2024, the global number of public charging points has already exceeded 5 million [7]. However, when many EVs charge simultaneously, particularly during

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 568–586, 2027.
https://doi.org/10.1007/978-3-032-30494-0_36

peak hours, local transformers and distribution lines risk overload. Upgrading grid infrastructure to accommodate these higher loads is expensive, complex, and time-consuming [18]. Urban areas face additional constraints, as retrofitting old buildings or underground parking facilities to accommodate charging equipment is complicated [2], and suitable parking spots are often scarce. Given these infrastructure limitations and the high cost of capacity expansion, operators must prioritize the intelligent use of existing resources: chargers, parking spots, and total power capacity. While power capacity allocation has received considerable attention in the literature, parking spot availability is often overlooked despite being equally constraining [19]. Safety regulations stipulate that an EV cannot be unplugged by anyone other than its owner once connected, meaning a vehicle may occupy both a charger and a parking spot for extended periods, even after charging is complete. This behavior is widespread in workplace or residential settings, where users leave their vehicles parked throughout the day [11]. Consequently, deploying too few chargers creates queues and user dissatisfaction, while deploying too many leads to low utilization rates and financial losses. Operators must therefore develop intelligent scheduling strategies to maximize the use of limited resources while preventing grid overload.

In this paper, parking spaces, chargers, and power grid capacity are treated as constrained resources. We consider a smart charging setting in which EV owners submit their charging demands in advance, and the station operator aims to maximize customer satisfaction by serving as many charging demands as possible. Studying the offline version of the problem is important as it provides a reference for evaluating online scheduling strategies and is also relevant in practical settings such as reservation-based charging infrastructures in airports and large cities. Under time-varying electricity prices, accepting more demands may increase electricity costs during peak-price periods, while prioritizing low-price periods may reduce the number of satisfied demands. Therefore, we formulate a bi-objective optimization problem that maximizes demand satisfaction while minimizing the total electricity charging cost of the accepted demands. We analyze how pricing profiles purchased from the grid utility influence charging scheduling decisions. The charging station resells electricity to EV users at a fixed price per kWh. The objective is to capture the trade-off between electricity procurement cost and demand satisfaction. By providing a Pareto front, our approach supports decision-making by allowing operators to determine how many charging requests to accept or reject under different pricing conditions, and to select a pricing strategy aligned with their desired level of service. This decouples the optimization model from any specific price-cost relationship, ensuring the two objectives remain structurally conflicting regardless of whether service prices exceed operational costs. Moreover, the introduction of heterogeneous chargers and the need to approximate a Pareto front significantly increase the computational complexity. To address this challenge, we propose and compare both an ε-constraint method and a Non-dominated Sorting Genetic Algorithm II (NSGA-II) enhanced with repair mechanisms and Pareto-front refinement to efficiently approximate the Pareto-optimal front.

The paper is organized as follows. Section 2 reviews the literature and positions our contributions. Section 3 describes the problem, while Sect. 4 presents the mathematical formulation and the ε-constraint method. Section 5 introduces the proposed NSGA-II. Section 6 reports the computational results. Finally, Sect. 7 concludes the paper.

2 Related Work

The EV Charging Scheduling Problem (EVCSP) determines when and where EVs should be charged under technical and operational constraints, in order to optimize given objectives. EVCSP formulations depend on optimization objectives that reflect stakeholder priorities (e.g., charging station or grid operators), the type of infrastructure (home, public, or private), and constraints related to EVs, charging facilities, and the power grid. Our analysis focuses on public and private charging stations, such as residential parking areas, shopping malls, campuses, and hospitals, while acknowledging that cross-study comparisons remain challenging due to heterogeneous operational contexts. In this context, existing research typically focuses on maximizing satisfied charging demand [22,24,25], reducing grid capacity usage [25], or minimize charging costs for users while still satisfying their demand [21]. In [13], a multi-objective formulation is considered, aiming to maximize both operator profit and customer satisfaction.

Below, we summarize the main constraints commonly encountered in EVCSPs, starting with EV arrival and departure times, which define the permissible charging window. An EV can only be charged within its plug-in period, i.e., from arrival to departure time [13,21,22,25]. The energy demand of each vehicle is typically defined either as a fixed quantity of energy (in kWh) to be delivered within the plug-in period [13,22,24,25], or as a flexible desired state-of-charge (SoC) interval [23]. In [16,17,22,24,25], the energy needs of each vehicle have to be fulfilled before departure. The number of available chargers and their technical characteristics, such as charging power rates, directly constrain the number of vehicles that can be charged simultaneously. While many studies assume homogeneous chargers [16,21,24], where all installed chargers have identical characteristics, other works consider non-identical chargers [13,22,23,25]. In these settings, particularly when the number of chargers is limited [16,21,22,24,25], charging schedules must prevent overlapping assignments on the same charger. Computational results reveal that deploying heterogeneous chargers (with varying output powers) is more advantageous in reducing the overall grid capacity requirement [25]. Chargers can follow constant power models [16,24,25], where EVs are charged at a fixed rate, or variable power models [12,21], which allow the charging rate to vary continuously. Preemptive scheduling, which allows a charging session to be interrupted and resumed at a later time, is considered in several studies [21–25]. In large EV charging stations, grid capacity constraints limit the simultaneous use of chargers due to the transformer's maximum power capacity, which caps the total power that can be drawn at any time [12,17,21,22,25]. Electricity prices are often represented using Time-of-Use tariffs, where predefined

time intervals reflect variations in grid demand and encourage charging during low-cost periods [17].

NSGA-II is one of the most widely used methods for multi-objective optimization in the EVCSP [5]. It consistently produces competitive Pareto fronts, closely approximating optimal solutions and outperforming several alternative algorithms [10,13].

Although multi-objective optimization is widely used in the EVCSP literature, most studies adopt either a power grid or a user-centric perspective, rather than that of charging station operators [5,13], particularly when addressing electricity cost minimization under time-varying prices. However, charging stations also purchase electricity from utilities and must balance operational costs with service levels. This work addresses this gap by integrating bi-objective optimization into a realistic charging station framework. Based on settings similar to those in [24,25], we extend the EVCSP model to account for heterogeneous chargers and charging efficiency, and introduce a bi-objective formulation that captures the marginal electricity cost of each additional accepted demand under time-varying prices. This formulation provides a decision-support tool for demand acceptance and pricing. Since real-world stations typically include chargers with different power ratings, this heterogeneity increases scheduling complexity by requiring joint assignment and timing decisions. To address this, we adopt a type-based formulation that reduces computational complexity. We further propose two solution approaches: an ε-constraint method and a NSGA-II algorithm enhanced with repair mechanisms and Pareto-front refinement. Finally, both methods are evaluated on instances with 10–400 EVs under four pricing scenarios.

3 Problem Description

We address the EVCSP in a charging station with m parking spaces, each equipped with a charger delivering constant power. Chargers are classified into k types according to their power rate, denoted by the set $K = \{1, \ldots, k\}$, where type $i \in K$ delivers power $w_i \in \mathbb{R}_{>0}$ (in kW), with $m_i \in \mathbb{N}$ chargers of type i available. Thus, $\sum_{i=1}^{k} m_i = m$. The station operates under a global power constraint $w_G \in \mathbb{R}_{>0}$ (in kW): at any time, the total power delivered by all active chargers cannot exceed w_G. We consider the realistic case where the total power that could be delivered if all chargers were active exceeds the grid capacity, i.e., $\sum_{i \in K} w_i \cdot m_i > w_G$. The scheduling horizon $H = \{1, \ldots, T\}$ is discretized into time slots of duration τ. The station operator provides charging services to customers who submit charging demands before arrival. Let $J = \{1, \ldots, n\}$ denote the set of charging demands. Each demand $j \in J$ is characterized by an arrival time $r_j \in \mathbb{N}$, a departure time $d_j \in \mathbb{N}$ (with $r_j < d_j$), and an energy requirement e_j (in kWh). Without loss of generality, we assume that arrival and departure times are expressed as integer time slot indices in H. The parking time slots for demand j are denoted by $H_j = \{r_j, \ldots, d_j - 1\}$, representing the discrete time intervals during which the EV is present at the station. Each demand can be

either accepted or rejected. An accepted demand is assigned to a charger and occupies it throughout the interval H_j, with energy demand e_j fully satisfied.

Charging is *preemptive*: it can be interrupted and resumed during the EV's plug-in time H_j. However, the EV remains parked and occupies the assigned charger throughout this interval. Let $\eta \in (0, 1]$ denote the charging efficiency of the station, defined as the ratio between the useful energy delivered to the EV battery and the electrical energy drawn from the grid. The charging time required to fulfill demand e_j using charger type i is $p_{ij} = \left\lceil \frac{e_j}{\eta w_i \cdot \tau} \right\rceil$, where p_{ij} is the number of required time slots.

The charging station purchases electricity from the power grid according to a time-varying tariff profile, where ep_t (€/kWh) denotes the unit price paid by the charging station at time slot $t \in H$. EV users are charged a fixed price per kWh by the station operator. Consequently, reducing electricity procurement costs increases the charging station's operational profit. The model schedules all charging demands to optimize two conflicting objectives: (i) Maximize the number of accepted charging demands to maximize user satisfaction and serve as many customers as possible, and (ii) Minimize the total cost of electricity paid by the charging station. Since these objectives are conflicting, we aim to approximate the Pareto front, i.e., the set of non-dominated solutions where no objective can be improved without degrading another. This enables operators to decide how many demands to accept or reject under a given pricing policy.

4 Mathematical Model Formulation

The original problem was formulated as a single-objective mathematical model in [22]. In the proposed bi-objective model, charger assignments are relaxed by considering chargers of the same type as identical, assigning EVs to charger types instead of specific chargers. To ensure parking feasibility, it is sufficient to guarantee that, at each time slot, the number of simultaneously used chargers of type i does not exceed the available number m_i. This relaxation reduces both the number of variables and the number of constraints, significantly lowering computational complexity. We introduce two sets of binary decision variables: $X = (x_{ij})$, where $x_{ij} = 1$ if charging demand j is assigned to charger type i (and 0 otherwise); and $Y = (y_{jt})$, where $y_{jt} = 1$ if demand j is charged during time slot t (and 0 otherwise). The bi-objective model then optimizes the following two objective functions:

$$f_1(X) = \sum_{j=1}^{n} \sum_{i=1}^{k} x_{ij}, \qquad f_2(X, Y) = \sum_{j=1}^{n} \sum_{t=r_j}^{d_j - 1} y_{jt}\, ep_t \sum_{i=1}^{k} x_{ij} \tau w_i. \tag{1}$$

Here, f_1 denotes the number of accepted charging demands, while $f_2(X, Y)$ represents the total electricity cost (€) paid by the charging station. To ensure the feasibility of any solution, the following constraints must hold:

$$\sum_{i=1}^{k} x_{ij} \leq 1, \qquad \forall j \in J, \tag{2}$$

$$\sum_{j=1}^{n} a_{jt} x_{ij} \leq m_i, \quad \forall i \in K,\ \forall t \in H, \tag{3}$$

$$\sum_{t=r_j}^{d_j-1} y_{jt} = \sum_{i=1}^{k} p_{ij} x_{ij}, \qquad \forall j \in J, \tag{4}$$

$$\sum_{j=1}^{n} \sum_{i=1}^{k} w_i x_{ij} y_{jt} \leq w_G, \quad \forall t \in H, \tag{5}$$

Each EV can be assigned to at most one charger type (constraints 2). Once assigned, an EV j continuously occupies a charger from its arrival r_j to its departure d_j, so the number of EVs using chargers of a given type at any time cannot exceed the available units (constraints 3). This is modeled through the parameters a_{jt}, equal to 1 if $t \in H_j$ and 0 otherwise, indicating whether EV j is present at time t. Each accepted demand must receive exactly its required energy e_j within its plug-in window (constraints 4). Finally, the total power consumption at any time must remain below the grid capacity w_G (constraints 5). Charging is allowed only within the EV's presence window, i.e., $y_{jt} = 0$ for all $t \notin H_j$. Given that the product of two binary variables is itself binary, the linearization is performed by introducing additional binary variables $Z = (z_{ijt})$ to represent the conjunctions $(x_{ij} y_{jt})$. The variable z_{ijt} equals 1 if and only if both $x_{ij} = 1$ and $y_{jt} = 1$. Accordingly, the second objective can be rewritten as: $f_2(Z) = \sum_{j=1}^{n} \sum_{t=r_j}^{d_j-1} \sum_{i=1}^{k} e p_t z_{ijt} \tau w_i$ and constraints (5) are replaced by the following linear constraints: $z_{ijt} \geq y_{jt} + x_{ij} - 1$, $z_{ijt} \leq y_{jt}$, and $z_{ijt} \leq x_{ij}$, $\forall i \in K$, $j \in J$, $t \in H$, together with $\sum_{i=1}^{k} \sum_{j=1}^{n} w_i z_{ijt} \leq w_G\ \forall t \in H$.

4.1 ε-Constraint Method

A Pareto front for the bi-objective problem can be computed using the ε-constraint method [6]. We first solve the single-objective problem max f_1 (ignoring f_2) to obtain $f_1^{\max}$, the maximum number of satisfied demands. Then, at each iteration ℓ, the mathematical model is solved by minimizing f_2 while adding the constraint $f_1(x) \geq \varepsilon^{(\ell)}$. The parameter $\varepsilon^{(\ell)}$ is updated at each iteration as: $\varepsilon^{(\ell)} = f_1^{\max} - (\ell - 1), \quad \ell = 1, \ldots, f_1^{\max} + 1$. Each iteration yields a Pareto-optimal solution. Since f_1 is integer-valued, a unit decrement is sufficient to enumerate all distinct Pareto-optimal points.

5 NSGA-II with Hybrid Repair and Pareto Refinement

The design of the evolutionary algorithm is based on the NSGA-II framework proposed by [3]. Classical genetic operators, namely crossover and mutation can produce infeasible solutions. Therefore, repair procedures are applied selectively: crossover violations are handled by a lightweight corrective procedure, since crossover is applied at higher frequency; a heavier ILP-based repair procedure is reserved for mutation. In addition, a periodic refinement phase is applied to the current Pareto front to recover rejected EVs and improve solution quality beyond what the genetic operators alone can achieve.

5.1 Solution Representation

We propose a type-based representation for the EVCSP, where a solution is represented as the set S of pairs $S = \{(\pi_j, \sigma_j) \mid j \in J\}$, where $\pi_j \in \{-1, 1, \ldots, k\}$ indicates the charger type assigned to EV j ($\pi_j = i$ if assigned to charger type i, and $\pi_j = -1$ if the EV is rejected). The vector $\sigma_j = (\sigma_{j,t})_{t \in H_j} \in \{0,1\}^{|H_j|}$ specifies the charging schedule of EV j over its parking time slots $H_j = \{r_j, \ldots, d_j - 1\}$, where $\sigma_{j,t} = 1$ if EV j is charging at time t, and 0 otherwise. This new representation offers a compact encoding by associating EVs with charger types rather than individual chargers, thereby reducing the search space. Since all chargers of type i are identical, any assignment remains feasible as long as the number of EVs charging on type i does not exceed m_i. w_{π_j} denotes the power rate of the charger type assigned to EV j. Let $J' = \{j \in J : \pi_j \neq -1\}$ be the set of accepted charging demands. Feasibility is assessed via aggregate measures: $U_{it} = |\{j \in J' : \pi_j = i \text{ and } t \in H_j\}|$ is the number of EVs assigned to type i that are present at time t, and $w_{Gt} = \sum_{\substack{j \in J' \\ t \in H_j}} w_{\pi_j} \sigma_{j,t}$ is the total power consumption at time t. The solution S is feasible if and only if:

$$U_{it} \leq m_i, \qquad \forall i \in K,\ \forall t \in H, \tag{6}$$

$$w_{Gt} \leq w_G, \qquad \forall t \in H, \tag{7}$$

$$\sum_{t \in H_j} \sigma_{j,t} = p_{\pi_j j}, \qquad \forall j \in J'. \tag{8}$$

Algorithm 1: ILP-NSGA-II

Input: N, G, α_c, α_m, R_{lp}, $I^{\text{lp}}_{\max}$, g_{ref}, T_{lp}, g_{start}
Output: Pareto front $\mathcal{F}_1$

Initialize, evaluate, and rank P_0;
for $g = 1$ **to** G **do**
  $Q_g \leftarrow \emptyset$;
  while $|Q_g| < N$ **do**
    Select p_1, p_2 by binary tournament from P_{g-1};
    if $\mathcal{U}(0,1) < \alpha_c$ **then**
      $(c_1, c_2) \leftarrow$TwoPointCrossover(p_1, p_2);
      foreach $c \in \{c_1, c_2\}$ *infeasible* **do** $c \leftarrow$WorstEVRejectionRepair(c)
    else $(c_1, c_2) \leftarrow (p_1, p_2)$;
    foreach $c \in \{c_1, c_2\}$ **do**
      if $\mathcal{U}(0,1) < \alpha_m$ **then** $c \leftarrow$Mutate(c)
      if c *infeasible* **then** $c \leftarrow$IterativeILPRepair$(c, I^{\text{lp}}_{\max}, R_{\text{lp}})$
      if c *infeasible* **then** $c \leftarrow$WorstEVRejectionRepair(c)
      Evaluate and add c to Q_g;
  $P_g \leftarrow$EnvironmentalSelection$(P_{g-1} \cup Q_g, N)$;
  if $g \geq g_{\text{start}}$ ***and*** $(g - g_{\text{start}}) \bmod g_{\text{ref}} = 0$ **then** RefineParetoFront(P_g, T_{lp})
if *refinement not applied at* $g = G$ **then** RefineParetoFront(P_G, T_{lp})
return $\mathcal{F}_1(P_G)$;

Since the EVCSP has conflicting objectives, we approximate the *Pareto front* of non-dominated solutions. A solution S_1 dominates S_2 if $f_1(S_1) \geq f_1(S_2)$ and $f_2(S_1) \leq f_2(S_2)$ with at least one strict inequality. We maintain a population $\mathcal{P}$ ranked by *fast non-dominated sorting* and diversified via *crowding distance* [3].

5.2 Framework of NSGA-II

Algorithm 1 presents the proposed NSGA-II, which evolves a population of N individuals over G generations. The initial population is generated using a constructive heuristic (Sect. 5.3). At each generation, offspring are produced via binary tournament selection and genetic operators (Sect. 5.4). Infeasible schedules are restored by selective repair procedures (Sect. 5.5), and a periodic refinement phase improves the current Pareto front (Sect. 5.6). Population diversity is maintained through fast non-dominated sorting and crowding distance [3]. All stochastic decisions are sampled from $\mathcal{U}(0, 1)$.

5.3 Initial Population Generation.

Before describing the initial population generation process, we first present the heuristic used to construct a single feasible solution.

Constructive Solution Heuristic. To construct a solution S, the heuristic begins by marking all EVs as rejected, i.e., $\pi_j = -1$ and $\boldsymbol{\sigma}_j = \mathbf{0}$ for all $j \in J$. A candidate set $J' \subseteq J$ of EVs to be considered for scheduling is then built according to one of four configurations (EVs in $J \setminus J'$ remain rejected): (1) all EVs in J, ordered by non-decreasing r_j; (2) all EVs in J, ordered by non-decreasing r_j, with ties broken by non-decreasing d_j; (3) all EVs in J in uniformly random order; and (4) a random subset of J in uniformly random order. Configurations (1)–(3) generally lead to solutions that accept a near-maximal number of EVs, thereby concentrating the initial population in the high-f_1 region of the Pareto front. Configuration (4) is therefore introduced to ensure that the initial Pareto front contains more than a single solution. Next, each EV $j \in J'$ is then scheduled using the heuristic described below, with charger types tried in increasing order of power rate. In configuration (3), this sequence is additionally shuffled uniformly at random before processing the EVs. If no charger type can successfully schedule the EV, then the EV remains rejected.

EV Scheduling Heuristic. Given a partial solution S, a charging demand j, a charger type i, a Boolean parameter `respectGridCapacity`, and a parameter `strategy`, the procedure attempts to assign EV j to charger type i and schedule its charging slots. It returns `true` if the assignment succeeds and `false` otherwise. The parameter `respectGridCapacity` determines whether the grid capacity constraint is respected (Eq. 7). If it is set to `false`, the procedure may generate intermediate solutions that violate the grid capacity constraint; such

solutions are later repaired by a correction procedure. The parameter `strategy` determines the time slot selection criterion for charging assignment.
Step 1: Parking Feasibility. Every slot $t \in H_j$ must have an available charger of type i, that is, $U_{it} < m_i$. Otherwise, the procedure returns `false`.
Step 2: Candidate-Slot Construction. The set of candidate slots $\mathcal{T}_{\text{cand}} \subseteq H_j$ is constructed as follows. If `respectGridCapacity = true`, only slots satisfying $w_{Gt} + w_i \leq w_G$ are considered. Otherwise, all slots in H_j are included in $\mathcal{T}_{\text{cand}}$. If $|\mathcal{T}_{\text{cand}}| < p_{ij}$, the procedure returns `false`.
Step 3: Slot Ordering. The candidate set $\mathcal{T}_{\text{cand}}$ is sorted according to the chosen `strategy`: (i) *Electricity Price-Based Ordering (EPO)*: slots are ordered by non-decreasing electricity price ep_t. (ii) *Grid Load-Aware Ordering (GLO)*: slots are ordered by non-decreasing grid load w_{Gt}, with ties broken by ep_t.
Step 4: Assignment. The EV is assigned to charger type i, i.e., $\pi_j \leftarrow i$. The first p_{ij} slots in the ordered set $\mathcal{T}_{\text{cand}}$ are selected. The indicators and counters are updated accordingly, namely $\sigma_{j,t}$ for the selected slots, as well as U_{it} and w_{Gt}.

ILP-Based Cost Minimization. The EV scheduling heuristic selects charging time slots greedily according to a predefined ordering criterion, which does not guarantee optimality with respect to electricity cost. Therefore, once charger type assignments $\{\pi_j\}_{j \in J'}$ are fixed, the remaining decision – selecting which time slots to charge each EV – can be formulated as an ILP and solved to optimality. Let S be the solution and $J' \subseteq J$ the set of accepted EVs with assigned charger types $\pi_j \neq -1$ for all $j \in J'$. Let $w_j = w_{\pi_j}$ denote the charging power rate assigned to EV j. With fixed charger assignments, we formulate a simplified ILP model to minimize total electricity cost by deciding only charging slots via binary variables y_{jt}: $\min f_2(Y) = \sum_{j \in J'} \sum_{t=r_j}^{d_j-1} y_{jt}\, ep_t\, \tau\, w_j$ subject to $\sum_{t=r_j}^{d_j-1} y_{jt} = p_j$, $\forall j \in J'$, $\sum_{j \in J'} w_j\, y_{jt} \leq w_G$, $\forall t \in H$, and $y_{jt} = 0$, $\forall j \in J'$, $t \notin H_j$. If the model is infeasible, the original solution S is returned.

Population Generation. The initial population is generated by running the constructive heuristic with varying parameter combinations to produce diverse individuals, all with `respectGridCapacity = true`. Two deterministic individuals are always generated first: configuration (2) with *GLO* followed by ILP-Based Cost Minimization (Sect. 5.3), and configuration (1) with *EPO* without ILP-Based Cost Minimization. The remaining $N - 2$ individuals are generated stochastically by sampling parameter combinations in fixed proportions 25% / 50% / 25%: configuration (3) with *GLO* followed by ILP-Based Cost Minimization; configuration (4) with *GLO* followed by ILP-Based Cost Minimization; configuration (4) with *EPO* without ILP-Based Cost Minimization. These combinations are shuffled before generation to avoid ordering bias in the initial population.

5.4 Generating Offspring

Two-Point Crossover. Given two parent solutions p_1 and p_2, two cut points $l_1 \leq l_2$ are drawn uniformly from $\{1, \ldots, n\}$. Offspring c_1 inherits the schedule (π_j, σ_j) of each EV j from p_1 for EVs outside the segment $j \notin [l_1, l_2]$ and from p_2 for EVs within; c_2 inherits complementarily. The Worst EV Rejection Repair procedure (Sect. 5.5) is applied to each infeasible offspring to restore feasibility.

Mutation. Mutation is applied with probability α_m. One operator is selected according to the probabilities α_{orig}, α_{cons}, and α_{swap}. Original mutation: a random EV is selected; if it is scheduled, it is removed and rescheduled on another charger type; otherwise, a scheduling attempt is performed. Destruct–Construct: $Q \sim \mathcal{U}(1, Q_{\text{des}})$ accepted EVs are removed, then up to $L \sim \mathcal{U}(1, L_{\text{cons}})$ rejected EVs are reinserted using the GLO scheduling strategy. Swap charger types: two charger types are randomly selected; their assigned EVs are unscheduled and reassigned to the opposite type. All operators invoke the EV Scheduling Heuristic (Sect. 5.3) with allowed grid capacity constraint violations (`respectGridCapacity` = `false`). The resulting infeasible offspring are subsequently repaired by Iterative ILP Repair (Sect. 5.5).

5.5 Feasibility Repair Procedures

Worst EV Rejection Repair. This lightweight procedure restores feasibility in two sequential phases. In the first phase, while the grid capacity constraint is violated, the EV that accumulates the highest total charging power across overloaded time slots is identified and rejected. This is repeated until the grid constraint is satisfied. In the second phase, while a charger capacity constraint is violated for some type k, the EV of type k that occupies the greatest number of time slots in excess of the available charger count is rejected. Both phases terminate as soon as the corresponding constraint is satisfied.

Iterative ILP Repair. This heavier repair procedure proceeds in two phases. In the first phase, while the grid constraint is violated and fewer than $I^{\text{lp}}_{\max}$ iterations have been performed, the EV contributing most to the violation is identified according to strategy $R_{\text{lp}} \in \{1, 2\}$: $R_{\text{lp}} = 1$ selects the EV with maximum energy density $e_j/(d_j - r_j)$; $R_{\text{lp}} = 2$ selects the EV with maximum cumulative contribution to grid overload. The selected EV is temporarily rejected, and an ILP-based cost minimization model (Sect. 5.3) is solved under the blended objective $\min(0.5, f_2 + 0.5, \tilde{w}G)$ within a time limit $T\text{grid}$. In this formulation, $\tilde{w}_G$ is introduced as a decision variable representing the grid capacity and is minimized as a secondary objective in order to restore feasibility. In the second phase, rejected EVs are sorted by increasing energy density and greedily reinserted while respecting the grid capacity w_G. If the solution remains infeasible, Worst EV Rejection Repair is applied as a final fallback.

5.6 Pareto Front Refinement

Due to its computational cost, refinement is performed periodically from g_{start} every g_{ref} generations, and at generation G if not previously applied. Rejected EVs are prioritized based on increasing energy density, with ties broken by decreasing temporal slack. Each rejected EV is processed sequentially, with at most 5 insertion attempts and up to 3 accepted variants per solution, under a time budget of 500 ms. Insertion relies on a cascade of operators: a randomized regret-k insertion ($k = 3$) first evaluates all charger types and selects randomly among the k best feasible assignments; if it fails, a single-ejection operator removes one conflicting EV (selected based on temporal overlap, energy density, and charging power), inserts the candidate, and attempts to reinsert the removed EV; if necessary, a multi-ejection operator (depth 2) removes multiple conflicting EVs and is accepted only if it strictly improves the number of satisfied EVs. A candidate is accepted only if it satisfies grid feasibility, strictly improves the number of satisfied EVs, and respects a cost tolerance $(1 + \varepsilon_{\text{cost}})$ with $\varepsilon_{\text{cost}} = 1.0$. Finally, an ILP-based cost minimization procedure is applied to the refined Pareto front within a time limit T_{lp}.

6 Computational Results

This section introduces the instances, details the parameter settings and tuning, and presents a comparative performance analysis.

6.1 Instances

EV Instance Generation. Synthetic benchmark instances are derived from the real-world EV charging transaction dataset of Zhang et al. [26], comprising 441,077 sessions collected across 13 stations in Jiaxing, China over a two-year period from late 2019 to 2021. After removing abnormal records, zero-energy sessions ($e_j \leq 0.5$ kWh), and duration outliers ($(d_j - r_j) \notin [5, 720]$ min), a hybrid non-parametric generation approach is applied. For each instance, 70% of charging demands are generated by a Gaussian kernel resampling of the empirical joint distribution $(r_j, d_j - r_j, e_j)$ [20], while the remaining 30% are drawn from a three-component Gaussian Mixture Model whose parameters are estimated directly from the cleaned data using the Expectation–Maximization algorithm [4,9]. This strategy is motivated by Kolmogorov–Smirnov tests that reject all standard parametric families (normal, log-normal, Weibull, gamma) for all three variables ($p < 0.001$), due to the multimodal structure of arrival times and the heavy-tailed distributions of charging duration and energy demand. To avoid trivial rejections and ensure that acceptance decisions are driven by resource constraints, generated charging sessions are required to be feasible, assuming a charger power of 22 kW. Therefore, any demand violating the condition is rejected and resampled. Seven instance sizes $n \in \{10, 20, 50, 100, 200, 300, 400\}$ are considered. For each size, 10 instances are generated, yielding 70 benchmark instances in total, indexed sequentially from 1 to 70. We consider three charger types $i \in \{1, 2, 3\}$

with power levels $w_1 = 11$ kW, $w_2 = 22$ kW, and $w_3 = 43$ kW. Instances 1–30 share the same settings ($w_G = 43$, $m_i = 1$, $m = 3$), while instances 31–40, 41–50, and 51–70 use $(w_G, m_i, m) = (114, 3, 9)$, $(190, 5, 15)$, and $(380, 10, 30)$, respectively. The grid capacity is defined as $w_G = \max(\max_i w_i,\ 0.5 \sum_i w_i)$, ensuring both feasibility with the fastest charger and a cap at half the total demand. All chargers operate with efficiency 0.92 [15]. The generated benchmark instances, descriptive statistics, and validation figures are publicly available at[1].

Electricity Prices Scenarios. Four electricity price scenarios were sourced from [8]. Their hourly tariffs over a 24-h horizon are shown in Fig. 1. Scenario 1 applies a flat two-level step tariff; Scenario 2 follows a smooth single-peak profile with a gradual morning increase and afternoon decrease; Scenario 3 exhibits continuously varying prices; and Scenario 4 is peak-heavy, with high prices dominating most of the day and only two short off-peak windows. Varying the price profile allows assessing the sensitivity of the scheduling algorithm to different tariff structures, and in particular its ability to shift charging load toward low-price periods under both simple and highly dynamic pricing patterns.

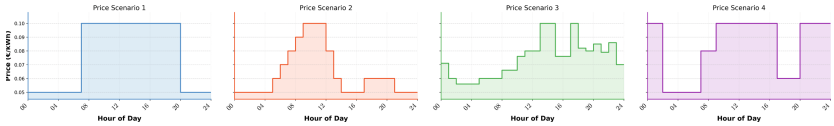


Fig. 1. price scenarios.

6.2 Parameter Settings and Tuning

The two multi-objective approaches described above are implemented: the ε-constraint method and the proposed NSGA-II, enhanced with multithreaded offspring generation. All algorithms are implemented in C++ and executed on a laptop (Intel Core Ultra 5 235U, 2.00 GHz, 16 GB RAM, Windows 11 64-bit). The time slot length is set to 6 min ($\tau = 0.1$ h). A time limit of 30 min per instance is imposed for the ε-constraint method, and the ILP models are solved using Gurobi 12.0.3.

NSGA-II remains sensitive to the selection of its parameter values. Therefore, we employed Optuna [1], a state-of-the-art Python framework for parameter optimization. The optimization process is guided by the hypervolume (HV) indicator defined in [27]. For each instance, we set the reference point for HV calculation as $r = \left(0,\ \max_t e_p^t \cdot \sum_j e_j\right)$ based on problem-specific bounds. We report the normalized hypervolume (NHV) in $[0, 1]$, where values closer to 1 indicate better Pareto front quality.

[1] https://github.com/imyzz/mo-evcsp-benchmarks.

Fig. 2. Parameter importance for small (left) and large (right) instances.

We use Optuna's TPE sampler. To account for instance size sensitivity, we conducted two separate analyses: one on three small instances (≤100 EVs) and another on three large instances (>100 EVs). Parameter importance was assessed via the absolute Pearson correlation between parameter values and NHV. For each configuration, the minimum NHV across the evaluated instances was maximized. To assess the benefit of refinement within NSGA-II, we introduced a boolean parameter `useRefine`: when set to true, the refinement is applied; otherwise, it is not. Parameter ranges and selected values are given in Table 1.

Table 1. "Parameter search space used in Optuna tuning.

Parameter	Type	Min	Max	Step	Small	Large	Parameter	Type	Min	Max	Step	Small	Large
N	Integer	10	100*	5	100	150	α_{orig}	Float	0.10	0.80	0.01	0.38	0.12
G	Integer	10	100	5	100	100	α_{cons}	Float	0.10	0.50	0.01	0.13	0.12
α_m	Float	0.01	0.30	0.01	0.18	0.18	α_{swap}	Float	0.10	0.60	0.01	0.17	0.21
α_c	Float	0.60	1	0.01	0.67	0.76	Q_{des}	Integer	1	15	1	1	13
L_{cons}	Integer	1	20	1	8	7	T_{lp}	Float	1	30	1	15	15
T_{grid}	Float	0.5	10	0.5	7.5	2.5	$I^{\text{lp}}_{\text{max}}$	Integer	1	10	1	6	6
R_{lp}	Categorical	1	2		2	2	`useRefine`	Boolean	True	False		True	True
g_{start}	Integer	0	G	1	44	90	g_{ref}	Integer	1	50	1	18	50

*Upper bound set to 150 for large instances ($n > 100$ EVs).

Results in Fig. 2 show a clear shift in sensitivity between instance sizes. For small instances, N, T_{Grid}, and g_{ref} are the most influential parameters. This is explained by the lower combinatorial complexity, where a larger population (N) improves Pareto front coverage, and frequent refinement (g_{ref}) enhances solution quality by reinserting rejected EVs and performing local improvements. The parameter T_{Grid} controls the time allocated to ILP-based repair used during mutation. Therefore, ILP repair has a significant influence, even though it is applied only in mutation, as it transforms infeasible solutions into feasible ones while minimizing electricity cost. For large instances, T_{grid} becomes the dominant parameter (≈ 0.30), followed by $\alpha_{Q_{\text{des}}}$ and `useRefine`. Since mutation

operators deliberately allow temporary infeasibility, ILP-based repair becomes critical, and insufficient time allocation degrades solutions. The importance of $\alpha_{Q_{\text{des}}}$ highlights the need for greater diversification. Finally, the growing relevance of `useRefine` suggests that refinement procedures, although computationally expensive, are becoming essential for improving solution quality.

6.3 Comparative Performance Analysis

In this section, we compare the two proposed approaches using three key performance metrics: the NHV as described in Sect. 6.2; Pareto-front distribution, assessed through the spacing indicator as defined in [14]; and running time. For the ε-constraint method, a single deterministic run is performed per instance. The developed NSGA-II, hereafter referred to as ILP-NSGA-II, is stochastic and is therefore executed 10 times independently for each instance to ensure statistical robustness and stability.

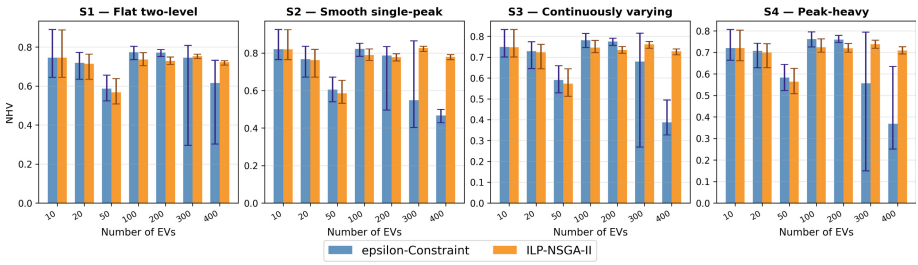


Fig. 3. NHV (mean, min, and max) per group and price scenario.

Figure 3 reports the mean NHV for each instance size group and price profile, with bars indicating the [min, max] range across the ten instances. For instances with 10 EVs, both methods achieve nearly identical performance (0.7590 vs 0.7584, i.e., −0.08%). For 20–50 EVs, ILP-NSGA-II performs slightly worse on average, with limited differences (−0.74% and −3.28%), a trend that continues for 100–200 EVs (−4.60% and −4.34%), although similar NHV values are reached in several runs. In contrast, for large instances (300–400 EVs), ILP-NSGA-II clearly outperforms the ε-constraint method, achieving gains of +21.71% and +59.89%, respectively, highlighting its superior scalability and stability. Note that results are reported as the average NHV over 10 runs, rather than the best value achieved by ILP-NSGA-II. An important observation concerns the influence of the price scenarios. ILP-NSGA-II shows consistent performance across all four scenarios, with limited variability. In contrast, the ε-constraint method is more sensitive to price variations. Scenario S1 (flat two-level) appears to be the easiest, yielding relatively high NHV values, whereas scenarios with more pronounced price variations (S2, S3, and especially S4) lead to lower NHV values, particularly for large instances.

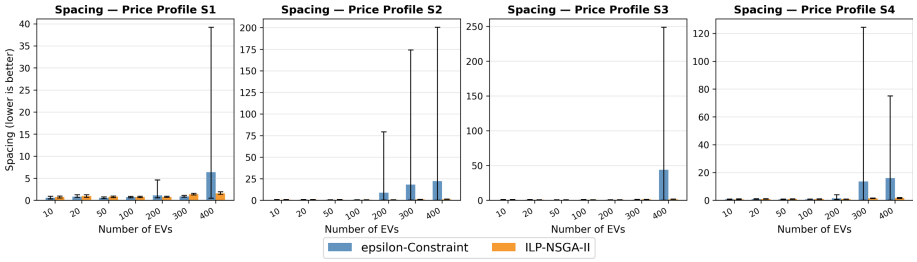


Fig. 4. Spacing (mean, min, and max) per group and price scenario.

Beyond NHV, Pareto-front quality is evaluated using the spacing indicator (lower is better), which measures the uniformity of solution distribution. Low spacing values indicate well-distributed non-dominated solutions. Results in Fig. 4 show that both methods achieve comparable spacing on smaller instances. As instance size increases, the spacing of the ε-constraint method rises more sharply and irregularly, reflecting a degradation in front quality due to time limits. In contrast, ILP-NSGA-II maintains more stable spacing values, demonstrating better robustness in preserving solution diversity.

Figure 5 reports the running time as a function of instance size for each price scenario. For instances with 10–50 EVs, both methods solve the problem very quickly (in less than 40 s). Starting from 100 EVs, differences begin to emerge: the ε-constraint method exhibits a sharp increase in CPU time. The gap becomes pronounced from 200 EVs onward. At this scale, the ε-constraint method requires more than 1000 s and reaches the imposed time limit (1800 s) for 300 and 400 EVs in most scenarios. ILP-NSGA-II maintains moderate execution times that grow smoothly with instance size and remains up to 77% faster, with an average reduction of 65% for instances ranging from 50 to 400 EVs. ILP-NSGA-II shows stable behaviour across all price scenarios, while the ε-constraint method is more sensitive to price structures, with lower execution times in S1 and significantly higher ones in scenarios with more pronounced price variations (S3 and S4). Both methods perform similarly up to 100 EVs, whereas ILP-NSGA-II becomes clearly preferable from around 200 EVs.

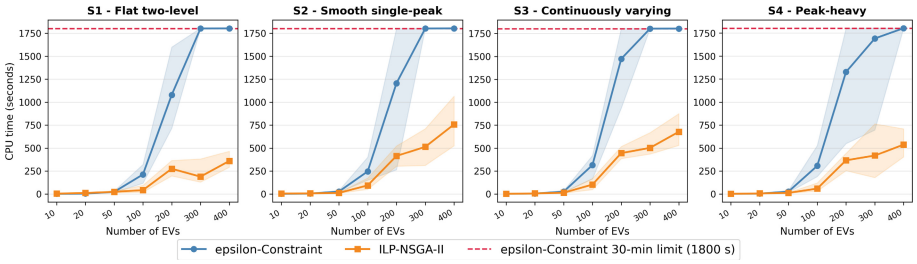


Fig. 5. Execution time per instance size.

Due to the large number of instances, only two representative cases are presented to illustrate the methods' Pareto front behaviour under different price scenarios. Figures 6 and 7 show Pareto front approximations for instances with 100 EVs and 400 EVs, respectively. The 100-EV instance is selected to represent a regime where the ε-constraint method slightly outperforms ILP-NSGA-II in terms of NHV, while the 400-EV instance illustrates a regime where the exact method collapses due to computational limitations, and ILP-NSGA-II provides a much better coverage of the Pareto front. Each colored dot corresponds to a non-dominated solution obtained from one of the ten ILP-NSGA-II runs, while blue stars represent solutions from the ε-constraint method.

For the 100-EV instance, both methods produce almost similar Pareto fronts across all price scenarios. However, differences appear in the low-satisfaction region (below approximately 40 satisfied EVs), where the ε-constraint method provides a denser and more complete coverage of the front, while ILP-NSGA-II yields fewer solutions. A clear impact of the price profile on the Pareto fronts can be observed. Scenario S2 consistently yields the lowest costs. Scenarios S1 and S3 exhibit similar cost ranges, leading to intermediate performance. In contrast, scenario S4 results in significantly higher costs.

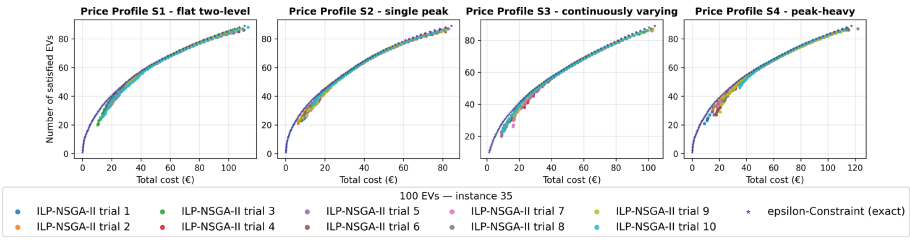


Fig. 6. Pareto front for instance 35 (100 EVs) across all price scenarios.

The 400-EV instance (Fig. 7) exhibits a markedly different behaviour. The ε-constraint front is partially truncated, especially under scenarios S2 and S4, due to the 30-minute time limit, resulting in sparse solutions confined to a limited region of the objective space. This is consistent with the near-1800 s CPU times

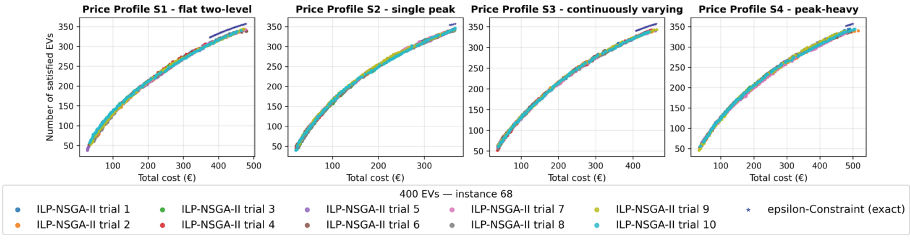


Fig. 7. Pareto front for instance 68 (400 EVs) across all price scenarios.

(Fig. 5) and the observed NHV degradation. In contrast, ILP-NSGA-II provides a dense and well-distributed approximation over a wider region. However, in the high-satisfaction region, its solutions remain slightly below those of ε-constraint, as the search is limited by the imposed computational budget (population size and number of generations). This suggests that additional search effort could further improve the approximation of extreme solutions. Consistent with previous observations, scenario S2 yields the lowest costs.

7 Conclusion

This paper studies a bi-objective EV charging scheduling problem with heterogeneous chargers and time-varying electricity prices, and compares an ε-constraint method with NSGA-II. Experiments conducted on 70 instances under four pricing scenarios show that the proposed NSGA-II significantly improves performance on large-scale problems, achieving up to +59.9% in normalized hypervolume for instances with 400 EVs (and +40.8% on average for instances with 300–400 EVs), while reducing computation time by 65% on average for instances with ≥ 50 EVs. Moreover, the method maintains well-distributed Pareto fronts. In addition, electricity price profiles are shown to significantly affect algorithm performance, particularly for exact methods. This conclusion holds for w_G set to 50% of the total charging power, with the understanding that lower values of w_G increase the problem's difficulty. Future work includes improving the ε-constraint method by avoiding unit-step enumeration, enhancing the NSGA-II framework, and conducting comparisons with other multi-objective algorithms. Additional directions involve exploring alternative charging station configurations and integrating more complex objective functions.

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

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019)
2. Crouse, M., Trousdale, D., Mulley, B., Mahmud, M., Cunningham, S.: Electric vehicle enabled buildings: evidence review of installation costs. Technical report, ClimateXChange (May 2022), report commissioned by the Scottish Government (2022)
3. 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)
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**(1), 1–38 (1977)
5. Elghanam, E., Abdelfatah, A., Hassan, M.S., Osman, A.H.: Optimization techniques in electric vehicle charging scheduling, routing and spatio-temporal demand coordination: a systematic review. IEEE Open J. Veh. Techno. **5**, 1294–1313 (2024)

6. Haimes, Y.: On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE Trans. Syst. Man Cybern. **3**, 296–297 (1971)
7. Global ev outlook 2025, International Energy Agency (2025). licence: CC BY 4.0
8. Karatzinis, G., et al.: Chargym: an EV charging station model for controller benchmarking. In: Maglogiannis, I., Iliadis, L., Macintyre, J., Cortez, P. (eds.) AIAI 2022. IFIPAICT, vol. 652, pp. 241–252. Springer. Cham (2022). https://doi.org/10.1007/978-3-031-08341-9_20
9. Lahariya, M., et al.: Synthetic data generator for electric vehicle charging sessions: modeling and evaluation using real-world data. Energies **13**(16), 4211 (2020)
10. Mahato, D., Aharwal, V.K., Sinha, A.: Multi-objective optimisation model and hybrid optimization algorithm for electric vehicle charge scheduling. J. Exp. Theoret. Artif. Intell. **36**(8), 1645–1667 (2024)
11. Mahlberg, J.A., Desai, J., Bullock, D.M.: Evaluation of electric vehicle charging usage and driver activity. World Electric Veh. J. **14**(11), 308 (2023)
12. Mathioudaki, A., Tsaousoglou, G., Varvarigos, E., Fotakis, D.: Online menu-based scheduling of electric vehicle charging. IEEE Trans. Smart Grid (2025)
13. Mishra, S., Mondal, A., Mondal, S.: A multi-objective optimization framework for electric vehicle charge scheduling with adaptable charging ports. IEEE Trans. Veh. Technol. **72**(5), 5702–5714 (2022)
14. Schott, J.R.: Fault tolerant design using single and multicriteria genetic algorithm optimization (1995)
15. Sevdari, K., Calearo, L., Bakken, B.H., Andersen, P.B., Marinelli, M.: Experimental validation of onboard electric vehicle chargers to improve the efficiency of smart charging operation. Sustain. Energy Technol. Assess. **60**, 103512 (2023)
16. Sone, S.P., Lehtomäki, J.J., Khan, Z., Umebayashi, K., Kim, K.S.: Robust EV scheduling in charging stations under uncertain demands and deadlines. IEEE Trans. Intell. Transp. Syst. (2024)
17. Tan, M., Ren, Y., Pan, R., Wang, L., Chen, J.: Fair and efficient electric vehicle charging scheduling optimization considering the maximum individual waiting time and operating cost. IEEE Trans. Veh. Technol. **72**(8), 9808–9820 (2023)
18. Tayri, A., Ma, X.: Grid impacts of electric vehicle charging: a review of challenges and mitigation strategies. Energies **18**(14), 3807 (2025)
19. Tran, C.Q.: Optimal deployment of charging infrastructures for urban networks (2022)
20. Van Kriekinge, G., et al.: Electric vehicle charging sessions generator based on clustered driver behaviors. Energies **14**(2), 37 (2023)
21. Wu, J., Su, H., Meng, J., Lin, M.: Electric vehicle charging scheduling considering infrastructure constraints. Energy **278**, 127806 (2023)
22. Zaidi, I., Oulamara, A., Idoumghar, L., Basset, M.: Maximizing the number of satisfied charging demands in electric vehicle charging scheduling problem. In: Legrand, P., et al. (eds.) EA 2022 LNCS, vol. 14091, pp. 89–102. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-42616-2_7
23. Zaidi, I., Oulamara, A., Idoumghar, L., Basset, M.: Electric vehicle charging scheduling problem: Heuristics and metaheuristic approaches. SN Comput. Sci. **4**(3), 283 (2023)
24. Zaidi, I., Oulamara, A., Idoumghar, L., Basset, M.: Maximizing the number of satisfied charging demands of electric vehicles on identical chargers. Omega **127**, 103106 (2024)

25. Zaidi, I., Oulamara, A., Idoumghar, L., Basset, M.: Minimizing grid capacity in preemptive electric vehicle charging orchestration: complexity, exact and heuristic approaches. Eur. J. Oper. Res. **312**(1), 22–37 (2024)
26. Zhang, Y., et al.: A high-resolution electric vehicle charging transaction dataset with multidimensional features in China. Sci. Data **12**, 643 (2025)
27. Zitzler, E., Thiele, L.: Multiobjective optimization using evolutionary algorithms—a comparative case study. In: Eiben, A.E., Bäck, T., Schoenauer, M., Schwefel, HP. (eds.) PPSN 1998. LNCS, vol. 1498 pp. 292–301. Springer, Heidelberg (1998). https://doi.org/10.1007/bfb0056872

PHD Showcase Papers

An Operating-System Infrastructure for Embedded BDI-Based Multi-agent Systems

Nilson Lazarin[1,2(✉)], Carlos Pantoja[1,3], and José Viterbo[2]

[1] Federal Center for Technological Education Celso Suckow da Fonseca (Cefet/RJ), Rio de Janeiro, Brazil
nilson.lazarin@cefet-rj.br, chon@grupo.cefet-rj.br
[2] Institute of Computing, Fluminense Federal University (UFF), Niterói, Brazil
[3] Linköping University (LiU), Linköping, Sweden

Abstract. The integration of Artificial Intelligence with Internet of Things devices has led to the emergence of Artificial Intelligence of Things, where intelligent decision-making capabilities are embedded directly in edge devices. Multi-Agent Systems (MAS), particularly those based on the Belief–Desire–Intention (BDI) model, provide a suitable paradigm for implementing autonomous behavior in such devices. However, deploying it on embedded platforms involves the computational overhead of the reasoning cycle and the complexity of the required software infrastructure. To this, this paper proposes an operating-system-level infrastructure for Embedded BDI-based MAS, introducing ACIRMa, an architectural model that extends the traditional embedded MAS architecture. Also, the paper presents an instantiation of the model through chonOS, a GNU/Linux-based operating system. Besides, an evaluation using a robotic prototype compares the specialized proposed OS against a general-purpose OS. Results show a 77.5% reduction in design time, and runtime experiments indicate improved efficiency, including a 14.4% reduction in memory consumption, a 46.8% reduction in swap usage, and a 7.2% increase in perception-processing capability.

1 Introduction

The convergence of Internet of Things (IoT) and Artificial Intelligence (AI) has led to the emergence of Artificial Intelligence of Things (AIoT), where intelligent decision-making capabilities are embedded directly into IoT devices. Since these devices generate large volumes of data, performing decision-making locally increases system autonomy while avoiding network congestion [16]. In this context, intelligent agents provide a natural approach for implementing autonomous decision-making. Agents are computational entities capable of perceiving their environment and selecting actions to achieve specific goals [39]. This paradigm has been widely adopted for developing systems capable of operating in complex environments and providing scalability, and flexibility [5].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 589–609, 2027.
https://doi.org/10.1007/978-3-032-30494-0_37

Agents often operate within Multi-Agent Systems (MAS), where multiple agents interact, exchange information, and coordinate their actions to achieve system objectives [8]. In AIoT scenarios, MAS are particularly advantageous due to their parallel and asynchronous nature, enabling distributed decision-making and the decomposition of complex problems into smaller responsibilities handled by individual agents [44]. Besides, autonomous agents typically exhibit characteristics such as *autonomy*, *reactivity*, *proactivity*, and *social ability* [8,46]. In addition, agents may exhibit *rationality* when they are expected to act in ways that promote the achievement of their goals rather than prevent them [46]. To support such capabilities, several cognitive models have been proposed for the design of intelligent agents [41].

Among these models, the Belief–Desire–Intention (BDI) architecture [7] is one of the most widely adopted approaches for constructing cognitive agents [11]. Over the past decades, the BDI model has driven the development of several programming languages, interpreters, and agent platforms [11]. One of the most influential BDI-based languages is AgentSpeak(L) [38], which provides a logical framework for programming agents using beliefs, events, and context-sensitive plans [12]. This language inspired several interpreters that bridge theoretical models of practical reasoning with executable agent systems [9].

Due to the advent of Single-Board Computers (SBC) in the 2010s, BDI-based agents can be directly situated in the physical world, perceiving and acting through sensors and actuators [21,31]. This capability enables the deployment of cognitive devices at the network edge. However, the strong abstraction provided by BDI programming, often relying on interpreters or intermediate representations, introduces challenges related to performance and resource consumption resource-constrained systems [1]. Moreover, these systems frequently depend on integration infrastructures for allow BDI-agent to interact with embedded firmware and physical devices [40].

These limitations represent open challenges for the adoption of BDI agents in AIoT environments. When successfully deployed in such contexts, BDI agents enable decentralized and autonomous decision-making directly at the network edge [26]. Providing a cognitive layer for embedded devices and mobile robots is particularly important for enabling rapid reactions to environmental changes while maintaining long-term objectives [40]. However, in the current state of the art, the computational cost of the BDI reasoning cycle and the complexity of the development and deployment process still hinder the practical adoption of embedded BDI-based MAS.

Thus, we formulate the following research problem: *in the current state of the art, the computational cost of the BDI reasoning cycle and the complexity of the development process hinder the deployment of BDI agents in embedded platforms.* To address this problem, this research investigates two questions: (i) *can the use of a dedicated operating system reduce the complexity of deploying BDI agents?* (ii) *is it possible to improve the performance of embedded BDI agents in resource-constrained platforms?*

This research is conducted following the Design Science Research (DSR) methodology [13]. To address the principle of *Problem Relevance*, Sect. 2 presents the theoretical background and a literature mapping that demonstrate the importance of the addressed problem and the lack of sufficient solutions in the current state of the art. To satisfy the *Design as a Search Process* principle, Sect. 3 discusses the rationale behind the proposed solution. Subsequently, two artifacts are introduced to address the identified problem: the ACIRMa architectural model (Sect. 3.1) and its instantiation through the *chonOS* operating system (Sect. 3.2). To satisfy the *Design Evaluation* principle, Sect. 4 demonstrates and evaluates the utility and feasibility of the developed artifacts. The implications of the results and their contributions to existing knowledge are discussed in Sect. 6, addressing the *Research Contributions* principle. Finally, the reproducibility of the research is presented in Sect. 5, ensuring the *Research Rigor* principle.

2 Background

Since the introduction of AgentSpeak(L) and its interpreters, several studies have explored the application of BDI-based agents across different domains. To understand how BDI-based embedded systems are currently being developed, we conducted a mapping study using the Snowballing method [45], as illustrated in Fig. 1. The work of Rao [38], which introduced this language, was used as the seed for the initial set of papers. In the *Forward Snowballing* step, citations of the *seed* paper were retrieved from the SCOPUS database, yielding 898 records as of October 2025. Only publications after 2010 were considered, resulting in 583 publications. To exclude works unrelated to embedded systems, we filtered titles or abstracts containing the terms "*robot**", "*IoT*", "*device*", "*physical*", "*embed**", or "*edge*", reducing the set to 155 publications. After removing duplicates, 63 candidate papers remained. Abstract reading identified 38 relevant publications presenting a *proof-of-concept* and a *model*, *approach*, *architecture*, or *tool*. After an initial full-text screening, 8 papers remained.

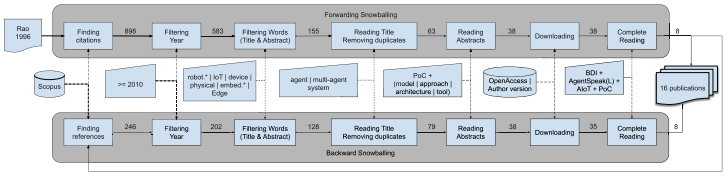


Fig. 1. Mapping review using Snowballing method.

In the *Backward Snowballing* step, references cited by these 8 papers were collected, resulting in 246 references. Applying the same year filter reduced this set to 202 publications. Filtering titles or abstracts with the same keywords reduced

the set to 128 papers. After removing duplicates and titles already considered, 79 unique papers remained. Abstract reading reduced this set to 38 candidates, of which 35 full texts were accessible. After an initial review, 8 additional papers were selected. Thus, the *snowballing* process resulted in 16 relevant papers which were analyzed in full, as summarized in Table 1.

Table 1. Publications analyzed in the literature review.

Paper	Framework	Arch/Infra	Middleware	Domain	Hardware
[20]	JaCa	CArtAgO	BrickPi	CPS	RPi 3
[27]	JaCaMo	EJasonArch	ROS	UAV	BeagleBone Black
[6]	JaCaMo	Physical Artifact	Javino	IoT	RPi Zero W
[10]	JaCaMo	Physical Artifact	Javino	IoT	RPi 3
[34]	Jason	ARGO	Javino	UGV	RPi Model B
[32]	Jason	ARGO	Javino	AmI	RPi Zero W
[19]	Jason	ARGO	Javino	UGV	RPi Zero W
[18]	Jason	Diffuser	Javino	V2V	RPi Zero W
[22]	Jason	Fuzzy-BDI	BrickPi	CPS	RPi 3
[35]	Jason	Interface Agent	RXTX	IoT	RPi Zero W
[21]	Jason	Jason Environment	BrickPi	CPS	RPi 3
[42]	Jason	Jason-ROS	ROS	UAV	RPi 3
[29]	Jason	Rason	ROS	UGV	Core i7 (laptop)
[31]	Jason	SAVI-ROS-BDI	ROS	UGV	RPi 4
[14]	Jason	SAVI-ROS-BDI	ROS	UGV	RPi 4
[43]	JasonEmbedded	ARGO	Javino	IoT	RPi Model B

The survey revealed the use of several BDI frameworks (Jason, JaCaMo, JaCa, and JasonEmbedded). All of these rely on the Multi-Agent Oriented Programming (MAOP) paradigm [4], which provides programming constructs that correspond to the abstractions used in a MAS to express social coordination.

Also, we can notice that, in general, cognitive embedded systems are composed of a *hardware platform* (a High-End IoT device) physically connected to several *hardware resources*, which may be provided by Low-End IoT devices. A Low-End IoT device is a system limited in computational resources (GPIO ports and serial connection), designed for basic sensing and actuation applications, and programmed with low-level firmware to manage sensors and actuators [30]. These devices cannot host a MAS but can be controlled by cognitive agents through serial communication protocols. Examples of such devices include microcontrollers from the ATMEGA or PIC families. A High-End IoT device, in turn, has sufficient computational resources (CPU, memory, and storage) to run a general-purpose operating system [30], hosting the MAS. These devices are single-board computers such as the Raspberry Pi or BeagleBoard.

Table 2. Embedded BDI-based systems Layers

Layer	Description
L4	*Reasoning Layer*. Hosts the cognitive agents, which receive percepts, deliberate, and send action commands.
L3	*Interfacing Layer*. Implements the communication protocol responsible for transmitting percepts and action commands.
L2	*Controlling Layer*. Resides in the microcontrollers, executing actions and translating raw sensor data into percepts interpretable by agents.
L1	*Actuating Layer*. Integrates sensors, actuators, and microcontrollers into the physical platform.

Table 3. Development Stages of embedded BDI-based systems

Step	Description
S1	*Physical Resources*. The designer selects sensors and actuators and builds the device according to the application domain.
S2	*Firmware Programming*. The designer implements functions to manage hardware resources and translate raw data into perceptions for agents.
S3	*Interfacing Programming*. The designer defines the communication protocol between the agents and the device hardware layers.
S4	*Agent Programming*. The designer implements the cognitive agents.
S5	*Services Management*. The designer configures the applications, services, and dependencies required by the embedded system OS.

Moreover, a generic architecture composed of four layers can be derived, as shown in Table 2. This four-layer view consolidates the MAOP dimensions into a practical stack for embedded autonomy: the Reasoning Layer captures AOP concerns; the Interfacing Layer materializes environment and communication abstractions and middleware; the Controlling Layer operationalizes device-level execution; and the Actuating Layer embodies the system within the physical world.

The development process of a MAS is similar to that of any other software system. Modeling languages, software engineering techniques, and different tools can support it. Essentially, the designer programs the MAS using an existing BDI framework tailored to their specific requirements. However, developing an autonomous and cognitive system embedded in devices is a little more complex. Considering the state-of-art, we notice five steps (S1-S5), as shown in Table 3, that a designer must overcome during the development process. Each of these steps requires a specific skill that must be employed along with possible support tools. Moreover, the following steps can be approached either in a top-down or bottom-up way, since the *reasoning layer* is loosely coupled to the *controlling layer* once they exchange messages.

Table 4. System Requirements

Req.	Description
R1	*Remote MAS Management*. Allows uploading, starting, and stopping the MAS remotely, avoiding physical access to the device.
R2	*Connectivity*. Enables automatic wireless network connection and network switching without compromising reasoning.
R3	*Network Zone*. Allows the device to identify its network location and connected devices after reconnection or network changes.
R4	*Wi-Fi Direct*. Provides point-to-point communication between the device and a computer when no wireless infrastructure is available.
R5	*Remote Firmware Management*. Supports *Over the Air* (OTA) firmware updates without wired connections.
R6	*BDI Compliance*. Supports hosting a BDI-based MAS as the reasoning mechanism.
R7	*Remote MAS Debugging*. Enables remote monitoring of agents' mental states, such as beliefs and intentions.
R8	*Hardware Interface*. Allows the MAS to interact with the physical environment through microcontroller management.
R9	*Openness*. Supports MAS openness, including agent mobility and communication with other MAS.

As S1 concerns mounting physical components, no automated tool supports it. Some tools can facilitate the process by offering integrated and automated solutions to help the designer in S2-S4, as presented by Jesus et al. [43]. Nevertheless, no solution was found to help the designer in S5. About the fifth step, we notice that almost all works have used Raspberry Pi or BeagleBone to host the MAS. This choice is justified because these *High-End IoT Devices* have the resources to support embedded technologies for AIoT [3]. For this type of device, there are *High-End IoT OS* options, which rely on different GNU/Linux-based distributions and even an IoT-oriented version of Windows® [2,17].

However, performance remains a challenge in embedded cognitive systems. Even when the hardware platform is capable, the responsiveness of the agent depends on the cumulative overhead introduced by perception processing, plan selection, inter-layer communication, and OS-level services. As demonstrated by Gavigan and Esfandiari [15], poorly structured BDI agents or tightly coupled reasoning mechanisms can significantly increase the duration of the reasoning cycle, reducing the system's ability to react to stimuli in time-sensitive contexts. Additionally, Mohajeri Parizi et al. [28] show that the execution architecture of agent frameworks—whether interpreter-based, compiled, or actor-based—substantially impacts concurrency, scalability, and throughput, reinforcing the need for lightweight reasoning pipelines in embedded scenarios. Likewise, experiments conducted by Amaral [1] indicate that CPU load, memory consumption, and latency become critical when these systems operate in AIoT scenarios.

Beyond performance constraints, as shown by Liu et al. [25], when using a general-purpose OS, default configurations are not adequate for embedded AI

systems—such as BDI-based MAS. According to Zhang and Huang [47], properly configuring a general-purpose OS is a complex task that demands domain knowledge and experience; for instance, the Linux kernel exposes more than 17 thousand configuration options for tuning. Consequently, adapting such systems to host embedded cognitive applications may significantly increase the complexity of the development process. This complexity also arises from the diversity of technologies involved and the breadth of knowledge required from the development team, since the development process spans five distinct steps—from assembling hardware resources to programming cognitive behavior using BDI abstractions.

Across these steps, we observe that any solution intended to support this workflow must address a set of desirable requirements related to hosting a BDI-based MAS in embedded devices. These requirements directly reflect the challenges faced during the development steps S1–S5, particularly those related to deployment, connectivity, and system configuration, which remain largely unsupported by existing MAS frameworks and middleware solutions. These requirements are listed in Table 4.

In the literature, efforts are concentrated primarily on providing middleware and frameworks that assist in steps 3 and 4, while the remaining steps—equally important for building autonomous and cognitive devices—remain largely unsupported. The retrieved technologies have indeed been used to develop embedded autonomous solutions with agents, and some have been extended to enable hardware interfacing and IoT communication. However, these works generally comply only partially with requirements R6–R8 and, in a few cases, R9—as none of them provides support for R1–R5 at the operating-system level.

For example, the use of ARGO [34], a Jason-extended architecture that interfaces microcontrollers, allowing BDI agents to access sensors and actuators by exchanging messages, has been used on different editions of the Raspberry Pi and on the Banana Pi, typically running Raspberry Pi OS, a general-purpose OS not dedicated to agent systems. With ARGO, the designer must still implement services to satisfy R1–R5, as the OS provides none of them. As reflected in Table 5, ARGO satisfies R6, R7, and R8, but does not provide openness (R9). To extend support for JaCaMo, ARGO was later repackaged as an importable module [24]. Similarly, the JasonEmbedded [33], in turn, significantly reduces this gap by integrating IoT communication, hardware interfaces, and MAS abstractions into a unified distribution derived from Jason. JasonEmbedded satisfies R6, R7, R8, and R9, although it still relies on Raspberry Pi OS and therefore lacks support for R1–R5.

Another relevant technology is the Robot Operating System (ROS) [37], a modular framework composed of tools, libraries, and algorithms organized as a communication layer running over a host operating system. Although widely used in robotics and compatible with MAS-based solutions, ROS supports S3, providing interfaces between robotic platforms and embedded software. It depends entirely on the host OS for services such as network management, debugging, and firmware updates, and thus satisfies none of the requirements

R1–R5. ROS does not provide native support for BDI agents, and consequently does not meet R6–R9. Nevertheless, works integrating ROS and Jason show that embedding BDI agents in robotic systems is feasible. For instance, JasonROS [42] and later applications [31] demonstrate that BDI reasoning can simplify robotic-agent development. These solutions satisfy R6–R8 but still lack agent mobility (R9) and do not address the OS-level requirements R1–R5.

As an alternative to Raspberry Pi OS, Windows 10 IoT is a Windows-based system prepared for IoT devices and offering built-in services that satisfy R1 and R3. Our research found no evidence of Windows IoT being used in BDI-based MAS, which is expected since the technology was not designed for agent systems. It does not support requirements R6–R9. However, due to its IoT-oriented features, Windows IoT remains a candidate worth exploring for intelligent autonomous applications using agents, provided that MAS mechanisms are implemented on top of it.

Despite the advances achieved by these technologies, the comparative analysis in Table 5 shows that none of the solutions simultaneously satisfies the performance demands and the development requirements (R1–R9). Current approaches either rely on general-purpose operating systems that hinder timely responsiveness, or depend on middleware and ad-hoc extensions that only partially address the complexity of hosting cognitive agents in constrained devices.

Table 5. Comparison of solutions to embedded BDI-based MAS.

Solution	R1	R2	R3	R4	R5	R6	R7	R8	R9
Windows 10 IoT	√	X	√	X	X	X	X	X	X
RPI	X	X	X	X	X	X	X	X	X
RPI + Jason	X	X	X	X	X	√	√	X	X
RPI + Jason + ARGO	X	X	X	X	X	√	√	√	X
RPI + JaCaMo	X	X	X	X	X	√	√	X	X
RPI + JaCaMo + ARGO	X	X	X	X	X	√	√	√	X
RPI + JasonEmbedded	X	X	X	X	X	√	√	√	√
RPI + Arduino	X	X	X	X	√	X	X	X	X
RPI + Arduino + JasonEmbedded	X	X	X	X	√	√	√	√	√
ROS	X	X	X	X	X	X	X	X	X
ROS + JasonROS	X	X	X	X	X	√	√	√	X

This reveals a clear gap: the absence of an operating-system level infrastructure specifically designed to support the deployment, management, and operation of embedded BDI-based multi-agent systems. As a consequence, designers are required to manually configure networking, firmware deployment, remote management, and debugging services, which significantly increases the complexity of developing and maintaining embedded cognitive devices.

3 A Specific-Purpose Operating System to Support Embedded BDI-Based MAS

The development of embedded BDI-based MAS involves challenges that extend beyond the design of the agents themselves. Designers must deal with heterogeneous hardware platforms, microcontroller programming, network configuration, and the deployment and monitoring of agents. In addition, executing such systems requires managing system-level services including connectivity, firmware updates, and debugging support, which significantly increase the complexity of the development process. In practice, most current solutions rely on general-purpose operating systems to host the MAS. However, these systems were not designed for BDI-agent-based applications. Consequently, several functionalities required for deploying these systems must be implemented manually, typically through ad-hoc scripts or external tools. This increases engineering effort, raises the likelihood of configuration errors, and makes deployments hard.

The mapping shows that solutions only partially address the requirements to support this class of systems. While some frameworks facilitate agent reasoning or hardware interfacing, other capabilities such as remote MAS management, automatic network configuration, firmware deployment, and debugging support remain unsupported at the operating-system level. Besides, from a design perspective, supporting autonomous embedded systems requires an infrastructure capable of handling device management, network connectivity, firmware deployment, and monitoring of agents while reducing the configuration burden placed on system designers. Considering these requirements, we extend the observed MAS architecture by introducing a specialized operating-system-level layer capable for managing system services required for the deployment and operation.

3.1 An Extension of the Embedded BDI-Based MAS Architecture

The architecture of embedded BDI-based MAS traditionally comprises four layers that connect hardware components to the reasoning system. In this work, we introduce an additional layer dedicated to system and network management. The resulting *ACIRMa* architecture therefore consists of five layers, as illustrated in Fig. 2, and is formally defined as:

$$ACIRMa = (D, F, P, L, E)$$

where:

D is the set of operating-system daemon services, where: $D = (d_1, d_2, ..., d_n)$;
F is a framework accountable for executing the cognitive layer (e.g., Jason, JaCaMo, JasonEmbedded);
P is a communication protocol or a middleware that enables interaction between the agents and the physical devices (e.g., Javino, ROS, BrickPi);
L is the set of low-end IoT devices, where $L = (l_1, l_2, ..., l_n)$;
E is the set of sensors or effectors, where: $E = (e_1, e_2, ..., e_n)$.

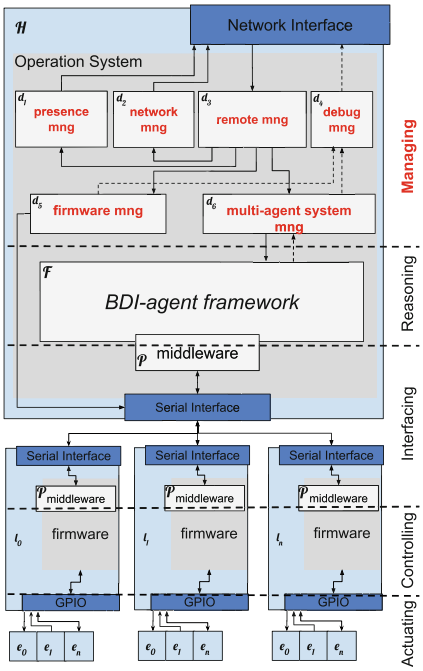


Fig. 2. ACIRMa architecture for embedded BDI-based MAS.

The lowest layer is the *Actuating Layer (L1)*, which includes the physical sensors and effectors (E_i) situated in the real-world environment. Each low-end device $l_i \in L$ is associated with a subset $E_i \subseteq E$ representing the sensors and/or effectors connected to that device, with the constraint that $E_i \neq \emptyset$, ensuring that every device (l_i) participates in either perception, actuation, or both. The next layer is the *Controlling Layer (L2)*, composed of the low-end IoT devices ($l_i \in L$, where $|L| \geq 1$) accountable for interacting with the physical components through firmware execution. Above it, the *Interfacing Layer (L3)* enables communication between the high-end IoT device (H) and the external hardware components. This layer relies on the protocol P to support firmware interaction with the connected L_i devices and to mediate the communication between the reasoning layer and the physical body of the system. The *Reasoning Layer (L4)* hosts the agent framework (F), accountable for executing the cognitive behavior of the system and managing the agents' beliefs, goals, and intentions.

Finally, the proposed *Managing Layer (L5)* provides management services (D) accountable for system configuration, network management, device administration, logging, and MAS lifecycle management. By integrating these capabilities at the operating-system level, the architecture simplifies the deployment and operation of embedded BDI-based MAS. To address the requirements R1–R9 identified in the previous section, the Managing Layer provides six operating-system services: d_1 a presence manager to solve R3; d_2 a network manager to

solve R2 and R4; d_3 a remote management to solve R1; d_4 a debug manager to solve R7; d_5 a firmware management to solve R5; d_6 a MAS manager to solve R6, R8, and R9.

It is important to emphasize that the novelty of the proposed approach does not lie in the individual management services themselves, since functionalities such as remote configuration, logging, or firmware deployment are already available in general IoT ecosystems. Instead, the contribution lies in integrating these capabilities as a unified operating-system-level infrastructure explicitly designed around the deployment, execution, management, and debugging requirements of embedded BDI-based systems identified in the literature review.

3.2 ChonOS Implementation

The proposed architecture is instantiated through the *Cognitive Hardware on Network - Operational System* (chonOS), a GNU/Linux distribution based on Debian. Debian was selected because it is widely used as the basis for other Linux systems, distributed under free software licensing, and supports several hardware architectures including ARM. It provides a stable software environment commonly adopted by single-board computers such as Raspberry Pi.

Table 6. Main services implemented in chonOS.

d_1 **chonos-ddns**: implements a *Dynamic DNS* mechanism accountable for maintaining the embedded system's *FQDN* updated, allowing access through a domain name without manually identifying its IP address when networks change. This module also implements a *Simple Service Discovery Protocol* server to announce the device within the local network, enabling multi-agent systems to exchange information about available services.
d_2 **chonos-network**: implements automatic management for *IEEE 802.3* and *IEEE 802.11* network interfaces, allowing direct access to the management layer through wired or wireless connections without manual configuration of *DHCP* . The module also supports *Wi-Fi Direct* and link monitoring with automatic reconnection using *ICMP* , which is useful for mobile robots operating outside wireless network range
d_3 **chonos-sysconfig**: provides a web interface for remote system management, extending the chonIDE [43], allowing configuration of device identification, network parameters, and interaction with the reasoning and firmware layers without requiring physical peripherals such as keyboard, mouse, or monitor.
d_4 **chonos-log**: provides remote monitoring through a web interface, allowing inspection of agent output messages and monitoring of agent behavior.
d_5 **chonos-firmware**: enables remote deployment and updates of firmware for Arduino boards without disassembling the cyber-physical system, using a firmware OTA mechanisms.
d_6 **chonos-embeddedMAS**: supports the deployment and execution of BDI-based MAS in the reasoning layer, enabling designers to upload, start, and stop projects developed using the JasonEmbedded [33] framework.

chonOS is ready-to-use, designed to support the deployment of embedded BDI-based MAS. It integrates several existing technologies and introduces specialized services to abstract technological dependencies involved in embedding MAS into hardware platforms. In addition, it integrates a framework for Embedded MAS programming, enabling agents to control hardware devices and supporting the development of mobile or IoT-based autonomous systems.

Since the system is designed specifically to support BDI-based MAS, minimizing operating-system overhead is essential to maximize the memory and processing resources available to the reasoning layer. The L5 was implemented in chonOS as a set of modular system services that together form its remote management infrastructure. These services support device configuration, MAS deployment, firmware management, and system monitoring. The main system services currently implemented in chonOS are summarized in Table 6.

4 Evaluation

According to Peffers et al. [36], DSR artifacts must be evaluated in relation to the requirements of their application context, providing evidence that the artifact effectively addresses the identified problem. To this end, we evaluate the instantiation of the proposed model from two perspectives. The first, *design-time evaluation*, compares the development and deployment process, while the second, *run-time evaluation*, analyzes resource consumption during execution. We aim to determine whether each approach (specific or general OS) simplifies the development and deployment of embedded BDI-based systems and whether it introduces additional computational overhead at runtime.

We consider the UGV domain by implementing an obstacle-avoidance prototype. The experiment compares the Cognitive Hardware on Network - Operating System (chonOS) (specific-purpose) and Raspberry Pi OS (rpiOS) (general-purpose). The latter was selected because the literature shows that most implementations rely on Raspberry Pi devices. The same hardware prototype, firmware, and BDI-agent code were used in both experiments in order to isolate the effect of the operating system.

The experimental prototype consists of an educational robotic platform, the chonBot, a low-cost 2WD prototype designed for robotic agents. The platform belongs to a family of standardized robotic prototypes composed of a single-board computer, a prototyping microcontroller, and a set of sensors and actuators, allowing BDI agents to interact with the real-world environment [23]. The hardware and software components of the prototype were organized as: L4 is hosted on a Raspberry Pi Zero W (RPi0W) (1 GHz single-core CPU, 512 MB RAM, and IEEE 802.11 b/g/n wireless LAN); L3 employs the Javino middleware; L2 is hosted on an Arduino Uno; and L1 includes an ultrasonic distance sensor, two reflective optical sensors, a light-dependent resistor, six light-emitting diodes, a buzzer, and the two-wheel mobile robot platform.

4.1 Evaluation 1: Development Complexity

This *design-time evaluation* analyzes how the chosen approach influences the development process. Specifically, we compare the tasks required to prepare the operating system, deploy the microcontroller firmware, and deploy the BDI-based reasoning system. These activities correspond to the S2–S5 stages of development described in the literature. We consider three aspects: the number of required configuration steps, the need for physical access to the device, and the degree of manual configuration required from the designer.

The general-purpose baseline was intentionally evaluated using the standard Raspberry Pi OS environment instead of lightweight command-line-only distributions. This choice reflects the deployment workflow commonly adopted in the related works, where designers rely on graphical interfaces, remote desktop access, external configuration tools, and manual system configuration to manage embedded MAS applications. Since one of the goals of this work is to reduce the operational complexity imposed on designers, the evaluation considers not only resource consumption, but also the practical deployment and management effort required during the development lifecycle of embedded BDI-based MAS.

Table 7 summarizes the tasks required to prepare the operating-system environment to host the embedded MAS (S5). While the specific-purpose approach centralizes configuration and deployment activities through a web-based management interface, the general-purpose approach requires physical access to the device and several manual configuration procedures, including network setup, installation of software dependencies, activation of remote-access services, and additional system configuration tasks. These differences reflect the operational complexity typically imposed on designers when deploying embedded BDI-based MAS using general-purpose operating systems.

Table 8 compares the process of deploying firmware to the microcontroller (S2-S3). With chonOS, firmware can be deployed remotely using the web interface. In contrast, the general-purpose approach requires the installation of additional tools (e.g., a VNC client) to allow the use of the GUI on the device.

Finally, Table 9 compares the tasks required to deploy and execute the MAS (S4). The specific-purpose platform provides a centralized environment for deployment and monitoring of the MAS. In the general-purpose platform, the developer must interact via VNC to execute the system and observe the output of agent messages, and must also use a web browser to visualize the agent mind.

The number of configuration steps corresponds to the individual actions performed by the designer to complete each stage described in Tables 7-9. Each step represents a single manual action required during the deployment process. Table 10 summarizes the main differences between both OS approaches.

Considering only the preparation of the operating-system environment, the general-purpose approach required 71 min, whereas the specific-purpose approach required only 16 min. This represents a reduction of approximately 77.5% in setup time. Together with the reduction from 16 to 5 manual configuration actions, these results indicate that the specialized OS substantially simplifies the

Table 7. Tasks required to prepare the OS environment.

OS Type	Configuration Tasks
specific-purpose	1) The designer powered on the prototype. The chonOS created a Wi-Fi Direct network named "EmbeddedMAS". The designer connected a computer to this network. 2) Using a web browser, the designer accessed the management interface (https://192.168.168.1:3270), assigned a *device name* and entered the wireless network credentials.
general-purpose	1) The designer needed to disassemble the prototype to connect a keyboard and a monitor to the device. After that the designer powered on the prototype. 2) The wireless network had to be manually configured. 3) SSH access had to be enabled to allow remote file transfer. 4) A VNC server had to be enabled to allow remote graphical access to the device. 5) The Java JDK had to be installed to support the BDI framework. 6) The Arduino IDE had to be installed to allow firmware deployment. 7) The Javino middleware and its dependencies had to be downloaded and installed. 8) The JasonEmbedded framework had to be downloaded from the project repository.

Table 8. Tasks required to deploy firmware to the ATmega microcontroller.

OS Type	Deployment Tasks
specific-purpose	1) The designer downloaded the source code from the repository to the computer. 2) Using the web management interface (*https://[deviceName].bot.chon.group:3270/*), the designer uploaded the project, compiled the firmware, and deployed it remotely.
general-purpose	1) The designer needed to connect a monitor to verify the IP address assigned to the device in order to access it remotely. 2) The designer accessed the device through SSH and downloaded the project repository. 3) The designer accessed the GUI using VNC to open the Arduino IDE. 4) The firmware was compiled and deployed to the microcontroller.

deployment preparation of embedded BDI-based MAS. Moreover, most tasks in the rpiOS required direct physical interaction with the device, while the chonOS allowed all operations to be executed remotely through the web-based manager.

Table 9. Tasks required to execute and monitor the MAS.

OS Type	Deployment Tasks
specific-purpose	1) Using the web management interface (*https://[deviceName].bot.chon.group:3270/*), the developer remotely starts and monitors the MAS project.
general-purpose	1) The designer needed to connect a monitor to verify the IP address assigned to the device in order to access it remotely. 2) The developer accessed the device using VNC to ensure that the MAS process would remain active even if the remote session was closed. 3) A terminal session was opened within the GUI to execute the MAS project. 4) On the computer, the designer used a browser to visualize the agent mind (*http://[deviceIPAddress]:3272/*).

Table 10. Summary of development tasks supported by each OS approach.

Task	specific-purpose	general-purpose
OS Installation	manual process	manual process
Network Configuration	remotely (web)	wired
MAS Runtime Dependencies	pre-installed	manual installation required
MAS Deployment	remotely (web)	remotely (VNC)
Debugging	remotely (web)	log terminal + remotely

An additional limitation observed in the general-purpose OS concerns the monitoring of agent minds. The JasonEmbedded framework exposes the internal reasoning logs through a web interface. However, due to the hardware constraints of the RPi0W (armv6l architecture), no browsers are available to run this interface. As a consequence, the debugging interface had to be accessed externally through the device IP address.

Finally, we notice that a specialized OS can reduce the configuration complexity required to deploy an embedded BDI-based MAS, positively answering RQ1.

4.2 Evaluation 2: Performance

The *run-time evaluation* analyzes the performance of the embedded BDI-based system during execution. Since such systems are typically deployed on resource-constrained devices, this experiment verifies whether the specialized OS introduces additional computational overhead during the execution of the BDI reasoning cycle. This evaluation focuses on analyzing the impact of the operating-system infrastructure on the execution of an embedded multi-agent system hosted locally on the device. Therefore, the experiment emphasizes the inter-

action between embedded hardware, inter-agent communication, and the BDI reasoning cycle during execution.

The experiment was executed using the same hardware platform described previously in order to isolate the effect of the operating system. Each benchmark was executed ten times for each OS. The reported values correspond to the average results obtained across these executions in order to mitigate possible variations caused by system scheduling and background processes. Both systems executed the same firmware and multi-agent program. To stress the reasoning cycle, we implemented a benchmark in which the Arduino microcontroller generated sequential numeric perceptions continuously during a 60-s interval.

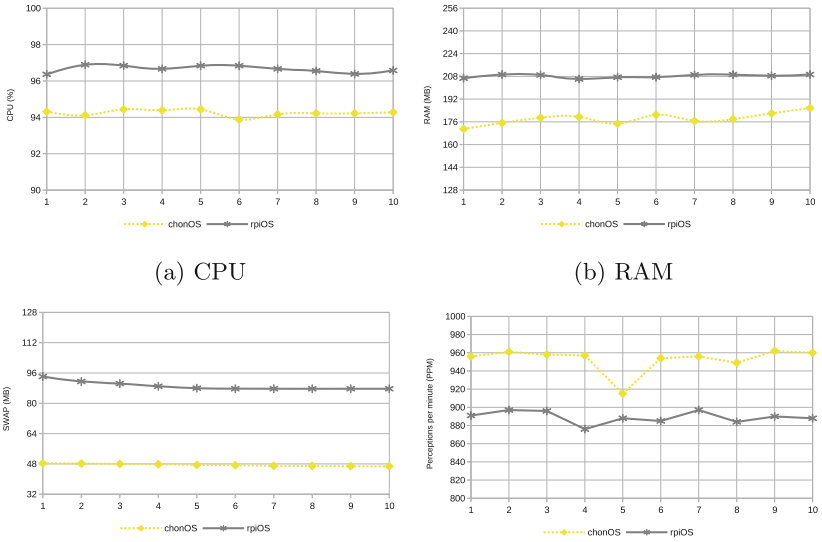
The benchmark employed two cooperative BDI agents within the MAS. The first agent was responsible for interacting with the embedded environment by receiving sequential numeric perceptions generated by the microcontroller. After receiving a perception, this agent forwarded the value through inter-agent communication to a second cognitive agent responsible for computing a Diffie–Hellman cryptographic key using modular exponentiation. Once the computation was completed, the result was returned to the percept-processing agent, allowing the next perception to be processed. This interaction created a continuous perception–reasoning cycle involving hardware interaction, inter-agent communication, and cognitive processing within the embedded MAS.

The total number of *processed perceptions during the one-minute interval* (PPM) was recorded as the system throughput. During execution, system resource consumption was monitored every five seconds, collecting CPU utilization, RAM usage, and swap activity. Figure 3 summarizes the measured results.

The experiment was executed ten times for each operating-system configuration. The reported values correspond to the mean and standard deviation of the observed measurements. The results indicate that both operating systems presented similar CPU utilization, with averages of $94.2 \pm 0.2\%$ for chonOS and $96.7 \pm 0.2\%$ for rpiOS. However, the specialized OS consumed less memory, averaging 178.2 ± 4.1 MB compared with 208.2 ± 1.2 MB in rpiOS, corresponding to a reduction of approximately 14.4%. Swap usage also presented a significant difference, with averages of 47.5 ± 0.6 MB in chonOS and 89.2 ± 2.2 MB in rpiOS, representing a reduction of approximately 46.8%. Regarding throughput, the MAS processed an average of 952.8 ± 13.8 PPM in chonOS and 889.2 ± 6.0 PPM in rpiOS, corresponding to an improvement of approximately 7.2%.

Overall, the results indicate that the specialized operating system does not introduce additional computational overhead when compared with the general-purpose environment. On the contrary, the reduction in memory and swap usage allows slightly higher processing throughput for the embedded MAS.

We notice that the use of a specialized OS does not negatively affect runtime performance and may even improve system efficiency, thus addressing RQ2.



(a) CPU (b) RAM

(c) SWAP (d) PPM

Fig. 3. Benchmark results comparing system resource usage during the execution of the embedded MAS.

5 Reproducibility

To facilitate replication of the research and provide transparency regarding the developed artifact and the evaluation process, we provide a dedicated page[1] containing detailed information about the experiments, source code, benchmark datasets, installation instructions for chonOS, videos demonstrating the prototype execution, and the data generated during the snowballing process.

6 Conclusion

The development of embedded BDI-based systems involves challenges that extend beyond agent programming. Designers must deal with heterogeneous hardware platforms, firmware development, network configuration, and system-level services required to deploy and maintain autonomous devices. As discussed in the literature, existing approaches rely primarily on general-purpose operating systems combined with middleware or framework extensions, which leaves several operational concerns unsupported and increases the complexity of the development process.

This work addressed this gap by proposing an operating-system-level infrastructure specifically designed to support this class of systems. Following the DSR

[1] https://papers.chon.group/ICCSA/chonOS/.

methodology, we identified the limitations of current solutions, derived a set of requirements (R1–R9), and designed an architectural model that extends the traditional embedded MAS architecture with a dedicated management layer. This model was instantiated through the implementation of *chonOS*, a GNU/Linux-based distribution that integrates services for network management, remote deployment, firmware updates, MAS lifecycle management, and system monitoring.

The evaluation considered both *design-time* and *run-time* perspectives. The first evaluation showed that the specialized operating system substantially reduces the configuration effort required to deploy embedded BDI-based MAS, decreasing the number of manual configuration actions from sixteen to five and reducing the environment preparation time by approximately 77.5%. This result indicates that integrating management services directly into the OS simplifies the development and deployment workflow. The second evaluation demonstrated that this specialization does not introduce additional computational overhead. On the contrary, the experimental results showed lower memory and swap consumption when compared with a general-purpose OS, while maintaining similar CPU utilization and achieving a slightly higher perception-processing throughput. These results provide evidence that the proposed artifact effectively supports the operational requirements of embedded MAS while preserving runtime efficiency.

The contribution of this work lies in the design of the ACIRMa model, an architectural extension that introduces a management layer for embedded BDI-based systems, and in its instantiation through chonOS, a specialized operating-system distribution that implements the proposed services. By integrating system management and cognitive agent execution within a unified infrastructure, the proposed model and its implementation reduce configuration complexity and support key operational requirements for embedded MAS. As a result, this work contributes to making the development and deployment of embedded BDI-based systems more accessible, reproducible, and manageable in practical scenarios.

Future work includes extending chonOS to support additional agent frameworks, exploring lightweight kernel configurations to further reduce system overhead, and evaluating the platform in larger multi-device deployments involving cooperative autonomous agents.

References

1. Amaral, C.J.: Embedding Multi-Agent System Frameworks (2018). https://periodicos.univali.br/index.php/acotb/article/view/12867
2. Antony, A., S., S.: A review on IoT operating systems. IJCA **176**(24) (2020). https://doi.org/10.5120/ijca2020920245
3. Bansal, S., Kumar, D.: IoT ecosystem: a survey on devices, gateways, operating systems, middleware and communication. Int. J. Wireless Inf. Netw. **27**(3), 340–364 (2020). https://doi.org/10.1007/s10776-020-00483-7
4. Boissier, O., Bordini, R.H., Hübner, J.F., Ricci, A., Santi, A.: Multi-agent oriented programming with JaCaMo. Sci. Comput. Programm. **78**(6) (2013). https://doi.org/10.1016/j.scico.2011.10.004

5. Bordini, R.H., El Fallah Seghrouchni, A., Hindriks, K., Logan, B., Ricci, A.: Agent programming in the cognitive era. Auton. Agent. Multi-Agent Syst. **34**(2), 1–31 (2020). https://doi.org/10.1007/s10458-020-09453-y
6. Brandão, F.C., Lima, M.A.T., Pantoja, C.E., Zahn, J., Viterbo, J.: Engineering approaches for programming agent-based IoT objects using the resource management architecture. Sensors **21**(23) (2021). https://doi.org/10.3390/s21238110
7. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bounded practical reasoning. Comput. Intell. **4**(3), 349–355 (1988). https://doi.org/10.1111/j.1467-8640.1988.tb00284.x
8. Burguillo, J.C.: Multi-agent Systems. In: Self-organizing Coalitions for Managing Complexity (2018). https://doi.org/10.1007/978-3-319-69898-4_5
9. BBădică, C., Budimac, Z., Burkhard, H.D., Ivanovic, M.: Software agents: Languages, tools, platforms. Comput. Sci. Inform. Syst. **8**(2) (2011) https://doi.org/10.2298/CSIS110214013B
10. Castro, L.F.S.d., Manoel, F.C.P.B., Souza de Jesus, V., Pantoja, C.E., Pinz Borges, A., Vaz Alves, G.: Integrating embedded multiagent systems with urban simulation tools and IoT applications. RITA **29**(1), 81–90 (2022). https://doi.org/10.22456/2175-2745.110837
11. De Silva, L., Meneguzzi, F., Logan, B.: BDI agent architectures: a survey. In: IJCAI'20 (2020). https://dl.acm.org/doi/abs/10.5555/3491440.3492124
12. D'Inverno, M., Luck, M.: Engineering AgentSpeak(L): a formal computational model. J. Log. Comput. **8**(3), 233–260 (1998). https://doi.org/10.1093/logcom/8.3.233
13. Dresch, A., Lacerda, D.P., Antunes, J.A.V.: Proposal for the conduct of design science research. In: Design Science Research, pp. 117–127 (2015). https://doi.org/10.1007/978-3-319-07374-3_6
14. Gavigan, P., Esfandiari, B.: Agent in a box: a framework for autonomous mobile robots with beliefs, desires, and intentions. Electronics **10**(17), 2136 (2021). https://doi.org/10.3390/electronics10172136
15. Gavigan, P., Esfandiari, B.: Quantifying the relationship between software design principles and performance in Jason: a case study with simulated mobile robots. Ann. Math. Artif. Intell. **92**(4), 775–795 (2024). https://doi.org/10.1007/s10472-023-09844-3
16. Hou, K.M., Diao, X., Shi, H., Ding, H., Zhou, H., De Vaulx, C.: Trends and challenges in AIoT/IIoT/IoT implementation. Sensors **23**(11), 5074 (2023). https://doi.org/10.3390/s23115074
17. Javed, F., Afzal, M.K., Sharif, M., Kim, B.S.: Internet of Things (IoT) operating systems support, networking technologies, applications, and challenges: a comparative review. IEEE Commun. Surv. Tutor. **20**(3) (2018). https://doi.org/10.1109/COMST.2018.2817685
18. Jesus, V.S.D., Lazarin, N.M., Pantoja, C.E., Manoel, F.C.P.B., Alves, G.V., Viterbo, J.: A middleware for providing communicability to Embedded MAS based on the lack of connectivity. Artif. Intell. Rev. **56**(S3), 2971–3001 (2023). https://doi.org/10.1007/s10462-023-10596-z
19. Jesus, V.S.d., Pantoja, C., Manoel, F., Alves, G., Viterbo, J., Bezerra, E.: Bio-Inspired protocols for embodied multi-agent systems. In: 13th ICAART (2021). https://doi.org/10.5220/0010257803120320
20. Karaduman, B., Tezel, B.T., Challenger, M.: Enhancing BDI agents using fuzzy logic for CPS and IoT interoperability using the JaCa platform. Symmetry **14**(7), 1447 (2022). https://doi.org/10.3390/sym14071447

21. Karaduman, B., Tezel, B.T., Challenger, M.: Rational software agents with the BDI reasoning model for Cyber–Physical Systems. Eng. Appl. Artif. Intell. **123** (2023). https://doi.org/10.1016/j.engappai.2023.106478
22. Karaduman, B., Tezel, B.T., Challenger, M.: On the impact of fuzzy-logic based BDI agent model for cyber–physical systems. Expert Syst. Appl. **238**, 122265 (2024). https://doi.org/10.1016/j.eswa.2023.122265
23. Lazarin, N.M., Pantoja, C.E., Viterbo, J.: Towards a toolkit for teaching AI supported by robotic-agents: proposal and first impressions. In: WEI 2023 (2023). https://doi.org/10.5753/wei.2023.229753
24. Lazarin, N.M., Pantoja, C.E., Viterbo, J.: Dealing with the unpredictability of physical resources in real-world multi-agent systems. In: ICAART 2023, vol. 14546. Cham (2024). https://doi.org/10.1007/978-3-031-55326-4_3
25. Liu, X., et al.: Operating systems for resource-adaptive intelligent software: challenges and opportunities. ACM Trans. Internet Technol. **21**(2) (2021). https://doi.org/10.1145/3425866
26. Mascardi, V., et. al: Engineering multi-agent systems: state of affairs and the road ahead. ACM SIGSOFT Softw. Eng. Notes **44**(1), 18–28 (Mar 2019). https://doi.org/10.1145/3310013.3322175
27. Menegol, M.S., Hübner, J.F., Becker, L.B.: Coordinated UAV search and rescue application with JaCaMo. In: PAAMS 2018, vol. 10978, pp. 335–338 (2018). https://doi.org/10.1007/978-3-319-94580-4_33
28. Mohajeri Parizi, M., Sileno, G., Van Engers, T., Klous, S.: Run, agent, run! architecture and benchmarking of actor-based agents. In: AGERE 2020. ACM (2020). https://doi.org/10.1145/3427760.3428339
29. Morais, M.G., Meneguzzi, F.R., Bordini, R.H., Amory, A.M.: Distributed fault diagnosis for multiple mobile robots using an agent programming language. In: ICAR 2015. IEEE (2015). https://doi.org/10.1109/ICAR.2015.7251486
30. Ojo, M.O., Giordano, S., Procissi, G., Seitanidis, I.N.: A review of low-end, middle-end, and high-end Iot devices. IEEE Access **6** (2018). https://doi.org/10.1109/ACCESS.2018.2879615
31. Onyedinma, C., Gavigan, P., Esfandiari, B.: Toward campus mail delivery using BDI. J. Sens. Actuator Netw. **9**(4), 56 (2020). https://doi.org/10.3390/jsan9040056
32. Pantoja, C., Soares, H.D., Viterbo, J., Seghrouchni, A.E.F.: An Architecture for the Development of Ambient Intelligence Systems Managed by Embedded Agents, pp. 215–249 (July 2018). https://doi.org/10.18293/SEKE2018-110
33. Pantoja, C.E., Jesus, V.S.D., Lazarin, N.M., Viterbo, J.: A spin-off version of Jason for Iot and embedded multi-agent systems. In: BRACIS 2023, vol. 14195. Cham (2023). https://doi.org/10.1007/978-3-031-45368-7_25
34. Pantoja, C.E., Stabile, M.F., Lazarin, N.M., Sichman, J.S.: ARGO: An extended jason architecture that facilitates embedded robotic agents programming. In: EMAS 2016 (2016). https://doi.org/10.1007/978-3-319-50983-9_8
35. Pantoja, C.E., Viterbo, J., Seghrouchni, A.E.F.: From thing to smart thing: towards an architecture for agent-based AmI Systems. In: KES-AMSTA 2019 (2020). https://doi.org/10.1007/978-981-13-8679-4_5
36. Peffers, K., Rothenberger, M., Tuunanen, T., Vaezi, R.: Design science research evaluation. In: DESRIST 2012. Springer (2012). https://doi.org/10.1007/978-3-642-29863-9_29
37. Quigley, M., et al.: ROS: an open-source robot operating system. In: ICRA 2009 (2009). https://scholar.google.com/scholar?cluster=143767492575573826

38. Rao, A.S.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: Van de Velde, W., Perram, J.W. (eds.) MAAMAW 1996. LNCS, vol. 1038, pp. 42–55. Springer, Heidelberg (1996). https://doi.org/10.1007/BFb0031845
39. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Always learning, Pearson, $3T^{\underline{O}}$ edn. (2016)
40. Santos, F., Hübner, J., Becker, L.: Design and analysis of a BDI agent model for route planning in a UAV. In: WESAAC 2015. SBC (2015). https://doi.org/10.5753/wesaac.2015.33313
41. Sichman, J.S., Demazeau, Y., Boissier, O.: When can knowledge-based systems be called agents? In: IX SBIA, vol. 9, pp. 172–185. SBC, Rio de Janeiro (1992). https://scholar.google.com.br/scholar?cluster=12989357293789094729
42. Silva, G.R., Becker, L.B., Hübner, J.F.: Embedded architecture composed of cognitive agents and ROS for programming intelligent robots. IFAC-PapersOnLine **53**(2), 10000–10005 (2020). https://doi.org/10.1016/j.ifacol.2020.12.2718
43. Souza De Jesus, V., Mori Lazarin, N., Pantoja, C.E., Vaz Alves, G., Ramos Alves De Lima, G., Viterbo, J.: An IDE to support the development of embedded multiagent systems. In: PAAMS 2023. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37616-0_29
44. Sycara, K.P.: Multiagent systems. AI Mag. **19**(2), 79–79 (1998). https://doi.org/10.1609/aimag.v19i2.1370
45. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: EASE '14, pp. 1–10. ACM, UK (2014). https://doi.org/10.1145/2601248.2601268
46. Wooldridge, M., Jennings, N.R.: Intelligent agents: theory and practice. Knowl. Eng. Rev.**10**(2) (1995). https://doi.org/10.1017/S0269888900008122
47. Zhang, Y., Huang, Y.: "Learned": Operating Systems. ACM SIGOPS Oper. Syst. Rev. **53**(1) (2019). https://doi.org/10.1145/3352020.3352027

DECODE: Overcoming the Perception Bottleneck in Small Vision-Language Models for Computer Science Education

Martina Ianaro(✉) and Maurizio Gabbrielli

University of Bologna, Bologna, Italy
{martina.ianaro3,maurizio.gabbrielli}@unibo.it

Abstract. Small Vision-Language Models (SVLMs) represent a critical frontier for Edge AI in Computer Science Education (CSEd), offering the efficiency and accessibility required for decentralized learning environments. However, the effectiveness of these models is currently undermined by a representational collapse of the vision encoder when processing high-fidelity technical diagrams, pseudocode snippets, and pointer representations. This paper introduces **DECODE**, a training-free framework designed to mitigate this *Perception Bottleneck* through Visual Structured Extraction (VSE), with direct application to university-level programming education. By decoupling visual perception from logical inference, VSE offloads the interpretation of instructional details onto a symbolic representation, allowing the student model to focus exclusively on reasoning. Our evaluation across diverse programming benchmarks reveals that the vast majority of model failures stem not from a lack of logical capacity, but from a fundamental failure in visual grounding. By implementing the DECODE protocol, we achieve a substantial parameter efficiency gain, where compact models match or exceed the reasoning performance of significantly larger architectures. This approach effectively reclaims the latent reasoning capacity of SVLMs without the computational overhead of fine-tuning. By providing a scalable and efficient path toward high-quality pedagogical support, DECODE represents a foundational step in democratizing AI tutoring across global STEM disciplines.

Keywords: Multimodal RAG · Small Vision-Language Models · Perception Bottleneck · Computer Science Education · Visual Semantic Extraction · Cross-modal Retrieval

1 Introduction

The rapid advancement of *Downscaling Intelligence* has catalyzed the development of SVLMs capable of running on consumer-grade hardware. In CSEd, these models offer a transformative opportunity for personalized, privacy-preserving

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 610–626, 2027.
https://doi.org/10.1007/978-3-032-30494-0_38

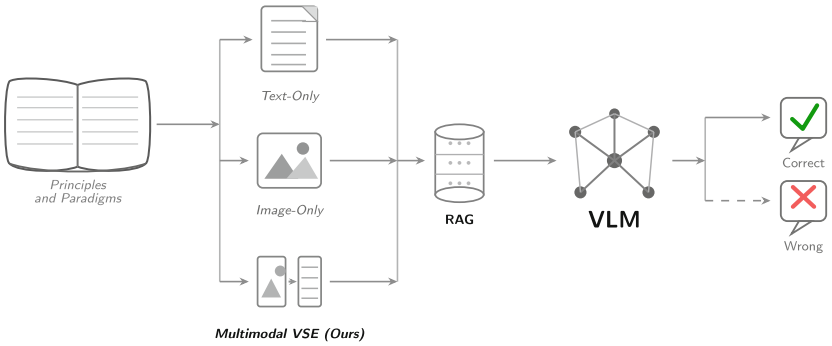


Fig. 1. Comparative architectural analysis of multimodal reasoning configurations. Starting from the textbook, the pipeline extracts multimodal data (Image-Only and Text-Only) alongside our Visual Structured Extraction (Multimodal VSE). Each modality is injected into the RAG knowledge base, which feeds the SVLM at inference time. Model outputs are assessed as correct or incorrect.

tutoring systems that can interpret both pedagogical text and technical illustrations [22,34]. However, model downscaling involves an unequal capacity regression with respect to the nature of the input: while the textual reasoning channel undergoes a moderate reduction, the visual channel suffers a drastic regression [7], ultimately inducing a *representational collapse* of the vision encoder. This phenomenon, which we term the Perception Bottleneck, is particularly consequential in CSEd, where diagrams such as memory stack and heap models, pointer chains, and finite automata are strictly relational and symbolic. According to *Dual Coding Theory* [4], mastering these abstractions requires the simultaneous processing of verbal and visual representations, such that a single perceptual error, for instance, misinterpreting a pointer arrow or a memory address label, can invalidate an entire logical inference chain, even when the model possesses the underlying domain knowledge to solve the problem. To address this, we present DECODE (*Decoupled Encoding of COntextual Diagram sEmantics*), a training-free implementation of the *Extract+Think* paradigm [7] designed to overcome the Perception Bottleneck in CSEd, see Fig. 1. Rather than forcing a constrained vision encoder to parse high-density technical symbols, DECODE employs a teacher model (GPT-4o) as an offline visual extraction oracle to transform raw diagrams into structured Instruction-Relevant Visual Details (IRVD), allowing the SVLM to focus its limited parameter count exclusively on the reasoning phase of the task. Our study utilizes a corpus derived from a foundational programming textbook [9], comprising 619 segments. We test our approach using PROG-VQA, a 1,000-item synthetic benchmark, and 50-VQA, a human-validated gold set designed to ensure ecological validity. Our results demonstrate that by isolating the perception loss, DECODE allows compact SVLMs to approach or exceed the performance of significantly larger archi-

tectures, confirming that the intelligence of small models is often masked by faulty perception. Our contributions are threefold:

1. We formalize the perception bottleneck in CSEd, demonstrating that CLIP-based retrieval [32] achieves an Image Retrieval Rate of 0.00 on formal diagrams, and that the vast majority of SVLM failures are perceptual rather than logical in nature.
2. We introduce DECODE, a scalable VSE pipeline that recovers cross-modal alignment without any model fine-tuning, offering a practical solution for resource-constrained educational environments.
3. We characterize a *Reasoning Threshold*, identifying the minimum model capacity required for successful perception–reasoning decoupling, and demonstrate that compact models can match or exceed the reasoning performance of significantly larger architectures.

2 Related Work

The Perception Bottleneck in Model Downscaling. A substantial body of work has investigated the failure modes of Vision-Language Models on tasks requiring visual grounding [3,23,33,40,43]. A recurring finding is the tendency of models to bypass visual content entirely, inferring answers from linguistic priors rather than genuinely interpreting the diagram [7,43]. Through a decoupled evaluation framework, Endo et al. [7] demonstrate that parameter reduction disproportionately degrades low-level perceptual capabilities relative to linguistic reasoning, ultimately inducing a *representational collapse* of the visual encoder. While this effect has been documented on scientific figures and mathematical diagrams [40,42], the behavior of SVLMs on formal programming diagrams remains an open empirical question that this work directly addresses. Crucially, Endo et al. [7] propose *visual extraction tuning* as a remediation strategy, which however requires additional supervised training on task-specific data. DECODE, by contrast, achieves the same decoupling in a training-free manner, substituting fine-tuning with an offline VSE protocol driven by a teacher model, making it directly deployable on resource-constrained hardware without any gradient update.

Symbolic Complexity and Dual Coding in CSEd. The perceptual bottleneck is particularly consequential in CSEd, where visual representations typically consist of technical diagrams, pointer chains, and memory layouts [9] whose semantics are relational and strictly domain-specific [4,6,35]. Dual Coding Theory [4] highlights the critical role of synchronizing verbal and visual representational channels, given that subject materials are deeply supported by instructional figures. Textual explanation alone is insufficient for students to construct accurate mental models of theoretical programming concepts [35]. This structural specificity renders general-purpose encoders like CLIP [32] ill-suited for the task, since a single perceptual error, such as misinterpreting a pointer arrow or a memory address label, invalidates the entire reasoning chain [23,40]. Zhang

et al. [40] further confirm this by showing that most multimodal models rely on textual priors rather than genuinely interpreting diagrams, with some models scoring higher *without* visual input at all. Unlike the mathematical diagrams studied in MathVerse [40], CS-specific diagrams such as pointer chains, memory maps, and abstract machines are non-redundant: no textual description accompanies them in the source material [9], making linguistic shortcutting impossible and visual grounding strictly necessary. Our empirical finding that CLIP-based retrieval achieves an Image Retrieval Rate of 0.00 on such diagrams provides direct quantitative support for this framing, and constitutes the core motivation for our VSE-based approach, which is specifically designed to operate on this class of non-redundant, strictly symbolic diagrams.

Multimodal Retrieval. The *Extract+Think* paradigm proposes decoupling perceptual grounding from logical inference as a principled response to the bottleneck described above [7,42]. Retrieval-Augmented Generation (RAG) has emerged as a dominant strategy for grounding model outputs in domain-specific knowledge without retraining [8,37–39], with a consistent finding across recent analyses [23,33] that textual descriptions of visual content outperform raw image embeddings in retrieval, suggesting that the challenge lies in how information is represented rather than in the generative capacity of the model [1,14]. Nowak et al. [23] confirm this across chart datasets but require a large vision encoder and task-specific training; Riedler and Langer [33] corroborate it in industrial settings, noting that image retrieval remains the primary failure point; Bharadwaj and Jain [1] further show that RAG benefits smaller models more than larger ones, a pattern we deliberately exploit through our reasoning threshold characterization. Unlike these approaches, which either operate on non-symbolic visual domains or require task-specific training and large vision encoders, DECODE addresses the perception bottleneck in the CSEd domain in a fully training-free manner, combining offline VSE extraction with a local, privacy-preserving inference stack that is directly deployable on resource-constrained hardware without any gradient update.

Ethics, Privacy, and the Case for Edge AI. The deployment of AI tutors raises critical concerns regarding student privacy and digital sovereignty [11–13,15,21]. Cloud-based inference routes sensitive educational data through third-party infrastructure [2,5], a concern that is particularly acute in low-resource settings where cloud reliance risks exacerbating the digital divide [10,17]. Edge AI and on-device inference represent a viable alternative [16,22,30,34]: by executing inference locally via OLLAMA [27], sessions remain private and fully functional without connectivity [34]. DECODE directly addresses this gap by enabling SVLMs to effectively process the complex formal diagrams essential to CSEd, providing high-quality pedagogical support that is simultaneously privacy-preserving and accessible on low-resource devices worldwide.

3 The PROG-VQA Dataset: Beyond Aesthetic Perception

The knowledge base in this project is built from the university course textbook *Programming Languages: Principles and Paradigms* [9]. The choice is strictly related to the goal of our work: building an educational AI tutor tailored to the students of the course, capable of answering with the specific terminology used and explained during lessons. To maintain the multimodal aspect, central in our work and in this domain, we extracted text and images from the book. The images are highly informative and technically focused.

3.1 The Challenge of CS-Specific Diagrams

The challenge lies in depicting these images as knowledge base entries. Here, the problem acquires the name of the perception bottleneck identified by [7]. The course concepts, apart from the visual component, could be covered by SVLMs due to their capacity to *Think*, which is not subject to the bottleneck. The problem remains limited to visual representation, which, if misinterpreted, entails a significant lack in the quality of answers, leaving them incomplete in terms of technical details. The *Extract* phase therefore suffers from *Representational Collapse* when faced with the high-density visual signals typical of CS formalisms.

3.2 Benchmark Composition

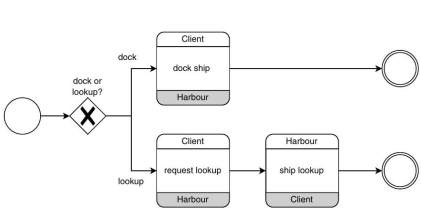
We curated the Prog-VQA benchmark, composed of two evaluation sets: the 50-VQA human-validated dataset, comprising 50 high-complexity diagrams extracted directly from the textbook [9] and manually annotated, and the Synthetic Augmentation Prog-VQA, composed of 1,000 visual question-answering pairs generated to ensure broad coverage of difficulty levels and diagram types.

4 Methodology: The DECODE Framework

The Decode framework is a training-free pipeline designed to mitigate the representational collapse of small vision encoders by decoupling multimodal perception from logical inference. Following the *Extract+Think* paradigm [7], the framework offloads the visual grounding of high-fidelity diagrams to a high-capacity teacher model, providing a clear symbolic representation to a compact student model and thereby bypassing the perception bottleneck that typically inhibits performance in Edge AI applications [22,34].

Data acquisition is structured into two distinct stages:

– *Stage 1: Structural Textual Segmentation.* Textual content is segmented using `PyPDFLoader` [18], preserving the logical structure of paragraphs alongside page-level metadata.

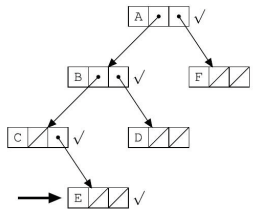


Q: How does the guarded choice work in this schema ?

✓ **DECODE:** In this abstract machine diagram, the guarded has to choose between `dock` or `lookupShip` where if the client selects `dock`, the interaction terminates; otherwise, `lookupShip` triggers a response.

× **Hallucination:** The guarded choice is a standard Erlang process management tool. It uses guard clauses to manage incoming requests in the VM.

Fails to consider the current image and visual branching structure in it.



Q: What is the value obtained by `*p`?

✓ **DECODE:** Looking at the memory map, `p` at `0x02` points to `0x01`. Since `x` is 10 at that address, dereferencing `p` results in **10**.

× **Perceptual Failure:** The pointer `p` seems to reference an unspecified location.

Fails to follow the pointer edge due to a lack in visual capability.

Fig. 2. Case studies on Programming Language Pedagogy visual queries. The **Baseline** (SVLM + CLIP) exhibits *Representational Collapse*, failing to resolve the relational structures of the abstract machine diagram (left) and the pointer chain (right). Our **DECODE** (SVLM + VSE) overcomes the *Perception Bottleneck* via the **Extract+Think** paradigm: on the left, it correctly identifies the branching logic; on the right, it successfully dereferences the pointer by reading the memory map.

– *Stage 2: Visual Element Isolation.* Visual extraction is performed using the `Unstructured` library [36], which isolates all figures present in the textbook and attaches metadata encoding their placement and positioning within the pedagogical flow.

The pipeline produced 619 segments, comprising 541 narrative text chunks and 78 visual elements. Although visual elements constitute only 12.6% of the corpus, they represent the semantic core of the domain, containing the high-density relational data, such as pointer addresses and memory layout structures, where SVLMs typically fail.

4.1 RAG System Architecture

We implemented a modular RAG architecture using a local Qdrant instance [31]. To isolate the impact of visual information on model performance, we constructed three distinct collections.

Decode—Multimodal VSE Collection. This collection implements our training-free extraction protocol. Unlike standard image captioning, which often produces generic descriptions, VSE is engineered to capture the specific formalisms of technical diagrams, including memory addresses, pointer relationships, and execution flows in code structures [23,33]. Visual descriptions were generated offline by GPT-4o acting as a zero-shot visual oracle, transforming raw pixels into structured VSE descriptors that serve as IRVD. As illustrated

in Fig. 2, symbolic relations are translated into text, for example representing a pointer p at address `0x02` pointing to x at `0x01`, allowing the SVLM to bypass its internal vision encoder entirely. Formally, the composite representation c_i for chunk i is:

$$c_i = \text{Enc}_{\text{MiniLM}}(d_i \,\|\, v_i), \quad c_i \in \mathbb{R}^{384} \tag{1}$$

where d_i is the text chunk, v_i the VSE descriptor, and $\|$ denotes concatenation. By providing the student model with a clear symbolic script of the diagram, we transform a complex multimodal reasoning task into a focused logical inference problem, allowing it to leverage its inherent linguistic and logical capabilities, which are disproportionately stronger than its visual perception [7,43].

Baseline Configurations. DECODE is evaluated against two alternative data representation strategies:

- *CLIP RAG (Native):* Visual content is embedded directly into the RAG system using *CLIP ViT-B/32* [32], with context retrieved based on visual similarity. This configuration maintains a purely multimodal architecture and serves as the primary reference point for quantifying the perceptual gap, as it represents the most common baseline in multimodal RAG literature [23,33].
- *Text-Only RAG:* Only the narrative textual segments from Stage 1 are encoded, excluding all visual metadata. This configuration isolates the model's internal linguistic priors, allowing us to distinguish failures due to missing visual grounding from those attributable to insufficient domain knowledge [1,8].

Because DECODE is entirely training-free, it circumvents the substantial computational costs and data requirements of fine-tuning or domain-specific pretraining, making it a scalable and accessible solution for deploying high-quality multimodal educational support on resource-constrained devices where privacy and low latency are critical [16,22,34].

5 Experimental Setup: Isolating the Perception Loss

The objective of our experimental design is to empirically isolate the perception bottleneck and quantify the extent to which representational collapse in vision encoders inhibits the reasoning potential of small-scale models. By adopting a training-free approach, we aim to demonstrate that compact Edge AI architectures can achieve performance parity with much larger models when visual grounding is properly structured through VSE [7,22,34].

5.1 Model Selection and Downscaling Strategy

To assess the scalability and architectural resilience of DECODE, we selected four representative SVLMs stratified by parameter count, examining how the

Table 1. Overall performance and efficiency metrics for human-graded and synthetic technical queries. The modalities comparison over models evaluated over our two benchmarks: Human Gold is the 50-VQA human validated benchmark and the 1K synthetic PROG-VQA benchmark. The names of the abbreviated metrics in the table are as follows: BS = BERTScore F1; R-L = ROUGE-L; Lat = avg. latency (s). † perception collapse of Moondream on visual items.

Strategy	Model	Human Gold (50)		PROG-VQA (1k)		Efficiency
		BS	R-L	BS	R-L	Lat
DECODE (Ours)	LLaVA-7B [24]	**.82**	**.25**	**.83**	**.31**	1.83
	LLaVA-Phi3 [25]	.79	.21	.79	.23	1.78
	Moondream† [26]	.55	.14	.66	.18	0.85
	Qwen2.5-7B [28]	.80	.25	.79	.23	4.40
CLIP RAG	LLaVA-7B [24]	.81	.24	.80	.24	1.64
	LLaVA-Phi3 [25]	.80	.22	.79	.21	1.72
	Moondream [26]	.79	.20	.73	.18	0.83
	Qwen2.5-7B [28]	.78	.21	.76	.17	3.15
Text-Only RAG	LLaVA-7B [24]	.82	.26	.82	.29	1.66
	LLaVA-Phi3 [25]	.81	.23	.80	.23	1.85
	Moondream [26]	.80	.21	.76	.21	0.82
	Qwen2.5-7B [28]	.80	.24	.78	.22	2.81

† Moondream: empty/nonsensical outputs on 34% (Human Gold) and 26% (PROG-VQA) of visual items under DECODE, deflating the average BS.

perception bottleneck manifests across the size spectrum. All models are served locally via OLLAMA [27], ensuring data sovereignty and reproducible latency, both of which are critical constraints for educational edge deployment [22,34].

- **Lower-Bound SVLMs:** MOONDREAM2 [26] (1.6B) serves as the extreme-efficiency lower bound, optimized for high-efficiency Edge AI deployment, while LLAVA-PHI3 [25] (3.8B) represents the *downscaling sweet spot*, combining mid-range multimodal reasoning with a compact footprint.
- **Upper-Bound References:** LLAVA V1.5 [24] (7B) and QWEN 2.5 [28] (7B) serve as high-capacity baselines to assess whether larger vision encoders (CLIP-ViT-L/14) or higher textual knowledge density can autonomously resolve the *representational collapse* without external grounding support.

The structured visual metadata provided by DECODE is hypothesized to enable the 1.6B model to approach the performance of the 7B baselines, validating perception-reasoning decoupling as a parameter-efficient strategy for Edge AI deployment.

5.2 Implementation Details

The entire inference stack runs locally to ensure data sovereignty and reproducibility [22,34]. All four VLMs are served via OLLAMA [27] on a single

Table 2. Ablative analysis performance over BERTScore F1 metric. The abbreviations in the table stand for: H50 for Human Gold 50-VQA and PV for PROG-VQA.

		By Difficulty						**By Type**			
		Easy		Medium		Hard		Visual (V)		Textual (T)	
Strategy	**Model**	H50	PV	H50	PV	H50	PV	H50	PV	H50	PV
DECODE	L7B	.83	.82	**.82**	**.83**	**.81**	**.82**	**.82**	**.79**	**.82**	**.84**
(Ours)	LP3	.83	.79	.78	.80	.79	.79	.79	.76	.81	.81
	M1.6‡	.80	.66	.48	.67	.57	.65	.52	.38	.79	.78
CLIP RAG	L7B	.82	.79	.82	.81	.80	.80	.81	.78	.80	.81
	LP3	.78	.77	.80	.80	.80	.79	.80	.78	.79	.79
	M1.6	.79	.67	.79	.74	.78	.76	.79	.63	.77	.77

L7B = LLaVA-7B [24]; LP3 = LLaVA-Phi3 [25]; M1.6 = Moondream2 [26].
‡ Perception collapse: Moondream performance drops under DECODE due to SVLM image-processing failures.

consumer-grade machine, without any cloud dependency.[1] A unified generation configuration was applied across all runs, with a near-deterministic decoding strategy and a fixed random seed to guarantee reproducibility; no model-specific prompt tuning was applied, so all differences in output are attributable solely to the retrieval strategy.[2] PDF ingestion follows the two-stage protocol described in Sect. 4: narrative text is segmented via *PyPDFLoader* [18] (LangChain), while figures are isolated with the *Unstructured* library [36], preserving page-level metadata and positional context. For the DECODE collection, VSE descriptors were generated *offline* by GPT-4o [29] as a zero-shot visual oracle, incurring no latency at inference time. The vector store is managed via a local QDRANT [31] instance. Responses are finally assessed using *BERTScore F1* [41][3] for semantic alignment and *ROUGE-L* for surface-level generation quality, as detailed in Sect. 5.5.

5.3 Comparative RAG Configurations

The core of our diagnostic strategy is a perception-reasoning ablation designed to detect representational collapse [7,43]. The three configurations isolate where the failure occurs:

1. **Native Multimodal RAG:** The SVLM receives raw pixels alongside the query, with the model's internal encoder responsible for parsing complex CS diagrams. We expect this to be the primary site of the bottleneck, as the integrated *Extract+Think* capacity is subject to representational collapse [23, 40].

[1] All experiments were conducted on a single NVIDIA GPU with 16 GB VRAM.
[2] Specifically: `temperature` = 0.1, `top_p` = 0.9, `num_predict` = 512 tokens, `seed` = 42.
[3] `distilbert-base-uncased`, via the `bert-score` Python library.

Table 3. In-depth Analysis: Error Distribution and Cross-domain Generalization. Part I: Error shift on the 50-item Gold Set between Native CLIP and Decode; Part II: Zero-shot performance across public benchmarks on the top line and between SVLMs architectures on the bottom line, comparing Image Only and DECODE.

Part I: Error Distribution (Human Gold Set)					
	Native CLIP		**Decode**		**Delta**
Error Category	Count	%	Count	%	
Perceptual Failure	38	76%	2	4%	−72%
Linguistic Hallucination	6	12%	3	6%	−6%
Reasoning Flaw	4	8%	2	4%	−4%
Total Failures	**48/50**	**96%**	**7/50**	**14%**	**+82% Acc.**
Part II: Zero-Shot Generalization and Scalability					
	Image Only		**Decode**		
Benchmark / Model	BS	Acc	**BS**	**Acc**	**ΔBS**
ChartQA [20]	.58	.03	**.64**	**.06**	+0.06
MMStar [3]	.63	.03	**.66**	**.11**	+0.03
ScienceQA [19]	.68	.04	**.69**	**.03**	+0.01
LLaVA-7B [24]	.60	.04	**.65**	**.10**	+0.05
Qwen2.5-7B [28]	.58	.01	**.62**	**.06**	+0.04
LLaVA-Phi3 [25]	.64	.11	.63	.08	(−0.01)

BS = BERTScore F1; Acc = Accuracy; ΔBS = Gain over Baseline.

2. **Text-Only RAG:** The model receives only narrative text segments, isolating its *pure reasoning* phase without any visual interference, thereby establishing the linguistic prior as a lower bound [1,8].
3. **Decode:** Following the protocol in [7], raw pixels are replaced with IRVD generated by the teacher model. By providing pre-extracted perception, we measure whether the SVLM can reclaim the reasoning performance of larger architectures, quantifying the *latent reasoning delta* achieved solely by resolving the perception bottleneck.

5.4 Cross-Domain Generalization Benchmarks

To evaluate the robustness of Decode beyond CSEd and address concerns about generalizability, we extend our analysis to three established multimodal benchmarks that require structured visual reasoning: ScienceQA [19], covering multimodal scientific reasoning where diagrams in the Natural Science domain require precise attribute extraction; ChartQA [20], involving complex logical reasoning over visual charts where SVLMs typically struggle with coordinate and value perception; and MMStar [3], specifically designed to probe whether vision-language models are genuinely interpreting visual inputs or merely exploiting

linguistic priors, directly addressing our central hypothesis on the perception bottleneck.

5.5 Evaluation Metrics and Methodology

ROUGE-L is computed to assess surface-level generation quality, while *BERTScore F1* [41] measures semantic alignment, penalizing hallucinations without penalizing correct paraphrases in open-ended VQA settings. As illustrated by the decoupling analysis in Fig. 4, the evaluation focuses on *Perceptual Fidelity*: the ability to correctly identify relational and symbolic elements such as pointer dereferencing. To isolate this, we specifically categorize instances of *Perceptual Failure* (see Table 3, Part I), defined as cases where the model's logical reasoning is sound but its visual input is corrupted by the perception bottleneck.

5.6 Ablation: Task Complexity and Visual Gap

We conduct a structured ablation along two dimensions. **Task Complexity:** Questions are categorized by the logical density of their associated visual data into *Easy*, *Medium*, and *Hard* levels. The bottleneck is hypothesized to be most pronounced on *Hard* queries, such as those involving multi-level pointer chains, where a single perceptual error cascades through the entire reasoning chain [23,40]. **Visual-Requirement Gap (V vs. T):** Questions that strictly require diagram analysis are labeled as visual queries (V), while the remainder are treated as textual (T). This division tests the decoupling hypothesis directly: SVLMs should remain competitive on textual queries regardless of the retrieval strategy, while degrading significantly on visual queries unless supported by DECODE.

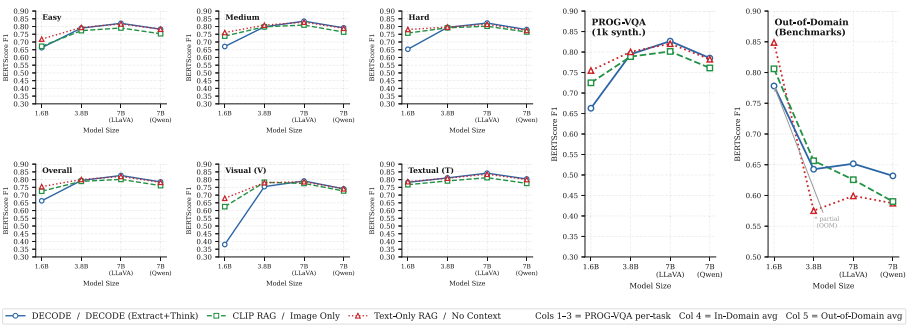


Fig. 3. Quantitative performance across task complexities and domains. Results are shown for three retrieval modalities across model sizes, under Task Complexity ablation (Easy/Medium/Hard) and Visual-Textual type split. The rightmost columns report Out-of-Domain generalization on CHARTQA [20], SCIENCEQA [19], and MMSTAR [3].

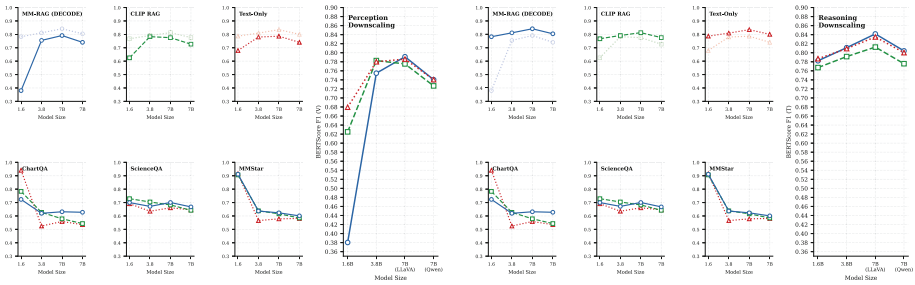


Fig. 4. The **Perception-Reasoning Asymmetry** in Model Downscaling. The left panel (*Perception*) evaluates performance on visually heavy queries only, while the right panel (*Reasoning*) isolates text-based queries, showing that logical capacity remains largely intact regardless of model size. Bottom rows report per-modality performance across out-of-domain benchmarks, confirming that the asymmetry between vision and logic is a domain-general phenomenon in technical visual reasoning.

6 Results and Discussion

As shown in Table 1, DECODE with LLAVA V1.5-7B achieves the highest overall performance, reaching a BERTScore of 0.82 on 50-VQA and 0.83 on PROG-VQA, outperforming all competing configurations. Notably, CLIP RAG consistently underperforms even the Text-Only RAG baseline when the underlying model is held constant, as observed for instance with QWEN 2.5-7B, confirming that raw image embeddings actively harm retrieval quality on formal CS diagrams. MOONDREAM (1.6B), the smallest model evaluated, degrades under DECODE due to information overload: the volume of structured visual details exceeds its reasoning capacity, placing it below the reasoning threshold identified in Sect. 1.

Table 2 provides a fine-grained breakdown by task complexity and query type across both benchmarks. DECODE with LLAVA V1.5-7B demonstrates strong robustness to increasing difficulty, maintaining a *BERTScore* in the range of 0.81–0.83 across Easy, Medium, and Hard queries, whereas CLIP RAG models exhibit a consistent performance drop as task complexity increases. The failure of MOONDREAM under DECODE on Visual-type queries is consistent with the information-overload hypothesis: the model cannot leverage the additional structured context, while CLIP RAG, which provides no symbolic detail, paradoxically avoids this collapse. LLAVA V1.5-7B with DECODE yields the strongest gains specifically on Visual (V) queries, where perceptual grounding is strictly necessary.

Table 3, Part I, reveals the magnitude of the perceptual failure under the Native CLIP baseline: 96% of responses are incorrect, with 76% of errors attributable to perceptual failure rather than reasoning deficits. Under DECODE, perceptual failures drop from 76% to 4%, yielding an overall accuracy gain of 82% points and confirming that the bottleneck is primarily perceptual, not cognitive. Table 3, Part II, reports zero-shot generalization performance on three

public benchmarks. Despite not being designed for these out-of-domain settings, DECODE yields positive Δ *BERTScore* on CHARTQA (+0.06), MMSTAR (+0.03), and SCIENCEQA (+0.01). These results are reflected in accuracy improvements, with scores increasing from 0.03 to 0.06 on CHARTQA and from 0.03 to 0.11 on MMSTAR. These results confirm that the VSE-based approach generalizes beyond CS-specific diagrams to technical visual reasoning more broadly, with the most consistent gains observed on mid-sized model architectures.

Figure 3 corroborates these findings quantitatively. The DECODE curve remains flat or improves as task difficulty increases from Easy to Hard, in contrast to both baselines, which degrade monotonically. Furthermore, the performance advantage of DECODE emerges consistently above a model size of approximately 3.8B parameters, which we identify as the reasoning threshold, below which, as in the case of MOONDREAM (1.6B), the additional structured context cannot be effectively exploited.

Figure 4 isolates the *downscaling asymmetry* directly: visual performance degrades sharply as model size decreases, while text-based reasoning remains largely stable across all model sizes. This asymmetry holds consistently across out-of-domain benchmarks, confirming that the gap between perceptual and reasoning capacity is a domain-general property of small vision-language models on technical visual content, rather than an artefact of the CSEd setting.

Qualitative Evidence. Figure 2 anchors these quantitative patterns in concrete failure modes. The Native CLIP baseline fails to follow pointer edges or resolve branching logic, whereas DECODE restores the Dual Coding synchronization necessary for correct mental model formation, effectively transforming a perceptually impaired SVLM into a grounded pedagogical tutor.

7 Conclusion

This paper has addressed the fundamental challenge of deploying SVLMs for CSEd, identifying a critical representational collapse within vision encoders as the primary barrier to effective multimodal pedagogy [7,43]. We have demonstrated that the failure of SVLMs to interpret technical diagrams is not a deficit of latent intelligence, but a failure of visual grounding: visual comprehension degrades catastrophically as models are scaled down, while reasoning capacity remains comparatively intact, confirming a pronounced *downscaling asymmetry*. Our analysis of the textbook-grounded corpus [9] revealed that while visual elements, such as memory models and pointer chains, comprise only 12.6% of the 619 extracted segments, they represent a critical semantic threshold where SVLMs consistently fail. The results from the PROG-VQA benchmark and the human-validated 50-VQA gold set demonstrate that offloading the interpretation of IRVD to a teacher model allows SVLMs to reclaim their latent reasoning capabilities, confirming that *Downscaling Intelligence* can be effectively managed in educational settings by bypassing faulty vision encoders [23,33]. Ultimately,

DECODE restores the pedagogical benefits of Dual Coding [4]: by providing high-fidelity textual descriptors of symbolic relations, it ensures that the verbal and visual channels remain synchronized, even within the constraints of a compact model. This is a crucial step toward the democratization of AI tutoring, ensuring that high-quality, privacy-preserving programming education remains accessible in resource-constrained environments worldwide [16,22,34].

8 Future Work and Limitations

Future work will extend DECODE along three directions: (i) autonomous, training-free extraction agents for multi-layered diagrams without predefined templates; (ii) validation across broader symbolic-heavy STEM disciplines to test the universality of the perception bottleneck [4,35]; and (iii) lightweight on-device vision-language adapters to eliminate the dependency on an external teacher model [16,34]. A further direction concerns evaluating the long-term pedagogical impact of structured visual grounding on students' understanding of abstract programming concepts [6,22].

The primary limitation of this study concerns its scope: the evaluation is grounded in a single programming textbook [9], restricting the direct generalizability of our findings to other curricula and instructional styles. A secondary limitation is the one-time dependency of the VSE phase on GPT-4o [29], which, while executed offline, is not yet fully consistent with a self-contained Edge AI deployment [22,34].

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

References

1. Bharadwaj, A., Jain, K.: Rag-assisted small language models for domain-level reasoning. In: 2025 Eighth International Conference on Image Information Processing (ICIIP), pp. 231–236 (2025). https://doi.org/10.1109/ICIIP68302.2025.11346385
2. Chaparadza, S., Bafana, S.: Promoting student sovereignty and data privacy in an ai-free higher education landscape: challenges and benefits. GAS J. Educ. and Lit. (GASJEL) **2**(9), 1–14 (2025)
3. Chen, L., et al.: Are we on the right way for evaluating large vision-language models? (2024). https://arxiv.org/abs/2403.20330
4. Clark, J., Paivio, A.: Dual coding theory and education. Educ. Psychol. Rev. **3**, 149–210 (09 1991). https://doi.org/10.1007/BF01320076
5. Deckker, D., Sumanasekara, S.: Student-centric ethical frameworks for ai-driven education: participatory consent and data ownership in a global perspective. Int. J. Sci. Res. Arch. **15**(3), 1491–1503 (2025)
6. Denny, P., MacNeil, S., Savelka, J., Porter, L., Luxton-Reilly, A.: Desirable characteristics for ai teaching assistants in programming education. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1., pp. 408–414. ITiCSE 2024, ACM (Jul 2024). https://doi.org/10.1145/3649217.3653574

7. Endo, M., Yeung-Levy, S.: Downscaling intelligence: exploring perception and reasoning bottlenecks in small multimodal models (2025). https://arxiv.org/abs/2511.17487
8. Fan, W., et al.: A survey on rag meeting LLMs: Towards retrieval-augmented large language models (2024). https://arxiv.org/abs/2405.06211
9. Gabbrielli, M., Martini, S.: Programming Languages: Principles and Paradigms. Undergraduate Topics in Computer Science, Springer (2010)
10. Gonzabay-Jiménez, E.F., Vargas-M, A., Velasteguí, H.J., Castro-Chuquiana, R.: Academic performance and inequality: Can generative ai widen the digital divide in higher education? In: International Conference on Emerging Technologies and Educational Innovation, pp. 193–204. Springer (2025)
11. Hamadeh, S., Amin, H.: AI, education and digital sovereignty. In: Frontiers in Education, vol. 10, p. 1677727. Frontiers Media SA (2025)
12. Holmes, W., Porayska-Pomsta, K.: The ethics of artificial intelligence in education. Lontoo: Routledge, pp. 621–653 (2023)
13. Huang, L.: Ethics of artificial intelligence in education: student privacy and data protection. Sci. Insights Educ. Front. **16**(2), 2577–2587 (2023)
14. Kazamel, S., Kocasarı, U., Alıcı, A.: An end-to-end computer vision system for structured information extraction from turkish id card images. In: Computational Science and Its Applications – ICCSA 2025, Part II, pp. 51–64. Springer-Verlag (2025). https://doi.org/10.1007/978-3-031-96997-3_4
15. Khan, W.N.: Ethical challenges of ai in education: balancing innovation with data privacy. AI EDIFY J. **1**(1), 1–13 (2024)
16. Koukaras, C., Stavrinides, S.G., Hatzikraniotis, E., Mitsiaki, M., Koukaras, P., Tjortjis, C.: Navigating the future of education: a review on telecommunications and ai technologies, ethical implications, and equity challenges. Telecom **7**(1), 2 (2026)
17. Landa-Blanco, M.: Artificial intelligence in education: applications and limitations for teachers in low-and middle-income countries. In: Frontiers in Education. vol. 10, p. 1681836. Frontiers Media SA (2025)
18. LangChain: Pypdfloader – langchain document loaders. https://docs.langchain.com/oss/python/integrations/document_loaders/pypdfloader (2024). Accessed 16 Mar 2026
19. Lu, P., et al.: Learn to explain: Multimodal reasoning via thought chains for science question answering (2022). https://arxiv.org/abs/2209.09513
20. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022). https://arxiv.org/abs/2203.10244
21. Moudgalya, S.K., Swaminathan, S.: Toward data sovereignty: Justice-oriented and community-based ai education. In: Proceedings of the 2024 on RESPECT Annual Conference, pp. 94–99 (2024)
22. Muminova, U., et al.: Real-time content personalization in educational IoT networks using on-device learning. J. Wirel. Mob. Netw., Ubiquit. Comput., Depend. Appl. **16**(2), 793–808 (2025)
23. Nowak, A., Piccinno, F., Altun, Y.: Multimodal chart retrieval: a comparison of text, table and image based approaches. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5488–5505 (2024)
24. Ollama Team: Llava on ollama library. https://ollama.com/library/llava (2024), 7B multimodal reference model

25. Ollama Team: Llava-phi3 on ollama library. https://ollama.com/library/llava-phi3 (2024), 3.8B parameter model based on Phi-3 Mini
26. Ollama Team: Moondream2 on ollama library. https://ollama.com/library/moondream (2024), ultra-light 1.6B parameter model
27. Ollama Team: Ollama: Run large language models locally. https://ollama.com (2024), framework for running LLMs locally
28. Ollama Team: Qwen2.5 on ollama library. https://ollama.com/library/qwen2.5 (2024), 7B reasoning-oriented model
29. OpenAI: Hello GPT-4o. https://openai.com/index/hello-gpt-4o/ (2024). Accessed 16 Mar 2026
30. Pinho, P.C., da Silva, R.Z., Primo, T.T., Dermeval, D., Bittencourt, I.I., Isotani, S.: Win-win situation: Generative ai in educational constrained environments. In: Simpósio Brasileiro de Informática na Educação (SBIE), pp. 1607–1616. SBC (2025)
31. Qdrant Team: Qdrant: High-performance vector similarity search engine. https://qdrant.tech (2024). Accessed 16 Mar 2026
32. Radford, A., et al.: Learning transferable visual models from natural language supervision (2021). https://arxiv.org/abs/2103.00020
33. Riedler, M., Langer, S.: Beyond text: Optimizing rag with multimodal inputs for industrial applications (2024). https://arxiv.org/abs/2410.21943
34. Shidaganti, G., Raj, V.A., Raman, V.M., Kumaran, S.: Edge computing in educational technology: The power of edge ai for dynamic and personalized learning. Edge of Intelligence: Exploring the Frontiers of AI at the Edge, pp. 121–152 (2025)
35. Sweller, J.: The worked example effect and human cognition. Learn. Instruct. **16**(2), 165–169 (2006). https://doi.org/10.1016/j.learninstruc.2006.02.005, https://www.sciencedirect.com/science/article/pii/S0959475206000193, recent Worked Examples Research: Managing Cognitive Load to Foster Learning and Transfer
36. Unstructured Technologies: Unstructured: Document processing for LLM applications. https://unstructured.io (2024). Accessed 16 Mar 2026
37. Xiong, P., Lacewell, C.W., Gobriel, S., Jain, N.: Context is all you need: Efficient retrieval augmented generation for domain specific AI. In: First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (2025). https://openreview.net/forum?id=6bKHoUQWlo
38. Xu, R., et al.: Simrag: self-improving retrieval-augmented generation for adapting large language models to specialized domains. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 11534–11550. Association for Computational Linguistics, Albuquerque, New Mexico (2025)
39. Yang, M., Ren, Y., Opoku, D.O., Li, R., Ren, P., Xing, C.: DSRAG: A domain-specific retrieval framework based on document-derived multimodal knowledge graph. arXiv preprint arXiv:2509.10467 (2025)
40. Zhang, R., et al.: Mathverse: does your multi-modal LLM truly see the diagrams in visual math problems? In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024, pp. 169–186. Springer, Cham (2025)
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTscore: Evaluating text generation with BERT (2020). https://arxiv.org/abs/1904.09675

42. Zhang, Z., et al.: Diagram formalization enhanced multi-modal geometry problem solver. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2025)
43. Zhou, Y., et al.: Scientists' first exam: Probing cognitive abilities of MLLM via perception, understanding, and reasoning. arXiv preprint arXiv:2506.10521 (2025)

Short Papers

A Machine Learning Approach for an End-to-End Music Inpainting Pipeline

Gustavo Romão[1], Elisa T. Albergaria[1], Diego R. C. Dias[2(✉)], Marcela Alves de Almeida[2], Vinícius F. S. Mota[2], and Leonardo Chaves Dutra Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ), São João del-Rei, Brazil
gugahr@aluno.ufsj.edu.br, {etuler,lcrocha}@ufsj.edu.br
[2] Universidade Federal do Espírito Santo (UFES), Vitória, Brazil
{diego.dias,marcela.almeida,vinicius.f.mota}@ufes.br

Abstract. Preserving musical memory is a constant challenge, as time is a key factor in the degradation of old scores. Over the years, stains, fading, and other physical damage erase notes and symbols, making reading difficult and jeopardizing the correct performance of these historical works. Among these symbols, musical clefs, reference markers that anchor pitch on specific staves, are particularly important, and as such, their absence makes the positions of subsequent notes semantically ambiguous. This study proposes a workflow that integrates an optical music recognition step based on YOLOv11 to detect clefs and note positions, statistical processing to convert the extracted staff position indices into compact meta-features of position distribution, and supervised classification for clef prediction. Using musical scores from the PRIMUS dataset, 950 images were selected and, according to three clef classes, the ability of the k-nearest neighbors, random forest, and gradient boosting models to predict missing clefs was compared, framing inpainting as a classification task. Gradient boosting showed the best performance, achieving an F1 score of 0.7242 on the test set. The results show that clefs can be predicted solely from positional statistics of pitch, which corroborates both the practical application of gap filling and an interpretable link between pitch patterns related to the register and the occurrence of clefs.

Keywords: Music Inpainting · Optical Music Recognition · YOLOv11 · Clef Prediction · Digital Preservation

1 Introduction

There are a number of original compositions confined to physical sheets of paper which are prone to continuous degradation. Stains, tears, ink fading, and chemical deterioration progressively compromise the legibility of notation, often leading to irreversible loss of musical content. Among the notational elements most

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 629–638, 2027.
https://doi.org/10.1007/978-3-032-30494-0_39

susceptible to such damage are musical clefs, foundational symbols that anchor the pitch reference of a staff. When clefs, in particular, become missing or illegible, the semantic interpretation of every subsequent note position is destabilized, rendering the correct reconstruction of the passage effectively ill-posed [1].

In order to mitigate this issue, musical restoration (or inpainting) has emerged as a research area that shifts the focus from the principle of visual restoration to the musical domain, allowing the reconstruction of missing elements while preserving coherence with the musical context. Previously, solutions to this problem depended on rigid rules based on musical theory. Today, the landscape has changed with the advancement of Artificial Intelligence, which allows models to learn musical styles and structures directly from data [4]. In this context, Optical Music Recognition (OMR) acts as an essential bridge, translating images of musical scores into a language that the computer can understand [5]. However, we still face a practical challenge: while most restoration research assumes that the data is already digitally organized, real-world preservation requires that we be able to read and extract information from worn images before even attempting any reconstruction.

In this work, we propose an approach that aims to address this gap by formulating a complete note-filling pipeline that integrates visual perception and predictive modeling. Specifically, we combine an optical clef recognition (OMR) stage based on the You Only Look Once (YOLOv11) algorithm to detect clefs and note positions in digitized scores. This stage consists of a statistical processing stage that converts the extracted pitch values into compact meta-features describing the pitch distributions, and a supervised classification stage that predicts the most plausible note, considering only these pitch-derived descriptors. In our experiments, we used musical scores from the PRIMUS dataset [2], with YOLOv11 trained to detect three classes of clefs and note positions with an input resolution of 640×640 pixels and a confidence threshold of 0.7. From the numerical representation of the resulting height (0-21), we constructed 10 descriptive statistical features per detected segment, resulting in a labeled dataset of 950 samples with naturally unbalanced class frequencies. We then compared three traditional machine learning families: k-nearest neighbors, random forest, and gradient boosting, using grid search fitting and 5-part stratified cross-validation, with the weighted F1 score as the main metric to account for class imbalance.

2 Theoretical Framework

In this section, we establish the theoretical framework necessary to bridge the gap between raw digitized scores and predictive modeling, beginning with an exploration of musical reconstruction as a data-driven reconstruction challenge. Thus, in this study we examine the role of musical record reconstruction (OMR) and the advantages of single-stage object detection, specifically YOLOv11, in extracting structured data from degraded manuscripts. Finally, we discuss the use of pitch distribution meta-features as compact and interpretable representations that capture the statistical signatures of musical register and tessitura.

Musical restoration can be understood as something similar to the task of visual restoration, since, given a partially corrupted artifact, the goal is to reconstruct the missing content while preserving coherence with the context [1]. In the musical domain, this problem is contextual, considering that the plausibility of a reconstructed element depends not only on the local notation but also on the structural regularities that govern the melodic contour, harmonic function, rhythm, and stylistic constraints. Initially, computational solutions attempted to repair sections of musical scores using manual rules based on traditional music theory. Over time, the field shifted direction: instead of programming each rule, we began using machine learning so that the models themselves could discover musical patterns directly from the data. Currently, research ranges from classical methods to cutting-edge technologies, such as generative models (VAEs) and Transformers, which can understand the structure of a piece of music in much longer and more complex contexts [4].

Within this context, clef reconstruction is a separate problem. Unlike a note or an accidental, which have their own meaning, the clef is what names the positions on the staff; without it, we don't know the actual pitch of any note that follows. Therefore, recovering a clef is, in fact, a work of deduction: we analyze the remaining notes to understand in which region of the scale the music fits. This makes the problem ideal for statistical methods that look at the distribution of notes (such as the mean and pitch variation), since these characteristics reveal the musical notation even when individual notes are blurred or damaged (Fig. 1).

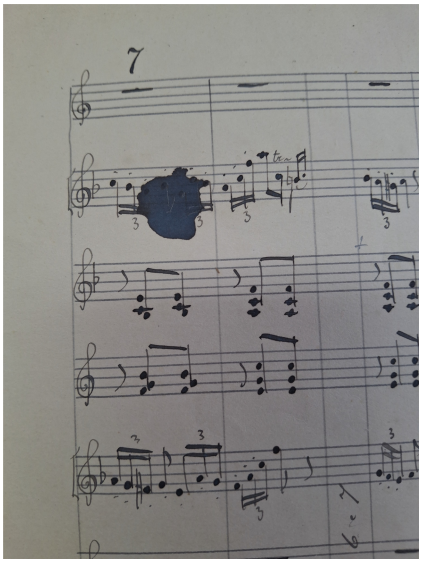


Fig. 1. An example of a music score with missing elements.

OMR is the tool that allows us to extract organized data from a musical score image. Previously, this process followed very rigid steps: it was necessary to erase the staff lines, isolate each symbol, and only then attempt to classify them one by one. Although it worked under ideal conditions, this method easily failed when faced with poor images or manuscripts worn by time. Today, the use of neural networks and object detectors has changed this scenario, allowing the computer to learn to identify symbols directly from the image, without needing those intermediate manual steps [6].

In this work, we use YOLOv11 as the visual basis of the system, treating musical notation as an object detection problem: clefs and notes are located and classified within identification frames. The efficiency of this type of technology for recognizing musical symbols has been proven in recent studies, showing that modern models are very robust at detecting small, close-up details, something common in musical scores [7]. Figure 2 illustrates YOLO being used to find and identify several musical symbols simultaneously.

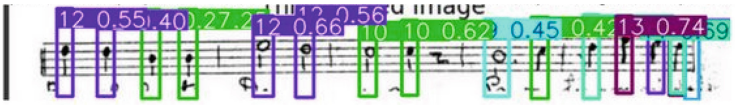


Fig. 2. YOLO applied to OMR: bounding-box detection of musical symbols (e.g., notes, clefs, accidentals) within a score.

Mapping symbolic music into informative numerical representations is a long-standing theme in computational musicology and Music Information Retrieval. Beyond note-by-note encodings, *distributional descriptors* such as mean, variance, skewness, and quantiles have been widely employed to characterize musical style, register, and authorship, capturing invariants that remain meaningful across notational variations.

Accordingly, this study paves way for clef inpainting through statistical meta-features extracted from OMR-derived positional values. Rather than modeling the full symbolic sequence, we represent each staff segment by compact distributional summaries (e.g., median, dispersion, and upper/lower quartiles). This design emphasizes interpretability through data, by making use of meta-features that have direct musical meaning (register, spread, asymmetry) and allow the learned decision function to be inspected in terms of musically plausible evidence. Prior work in mathematical and statistical perspectives on music supports the notion that pitch distributions encode discriminative signatures related to register and tessitura, enabling meaningful differentiation even without explicit symbolic reasoning about staff semantics [3].

Together, these foundations motivate an end-to-end formulation in which OMR extracts positional evidence from images, distributional meta-features summarize that evidence in a compact and interpretable form, and supervised learning performs clef inference as a classification task grounded in register-related statistics.

3 Methodology

We operationalize musical clef inpainting as a supervised classification task grounded on pitch evidence extracted directly from score images. The proposed architecture follows a sequential four-stage pipeline, as illustrated in Fig. 3.

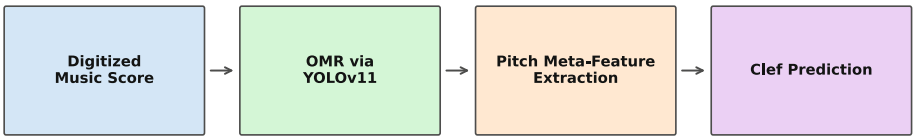


Fig. 3. Diagram illustrating the proposed end-to-end clef inpainting pipeline.

The Optical Music Recognition (OMR) stage is not treated as an external prerequisite, but as an integral component of the pipeline that converts unstructured visual notation into structured, machine-learning-ready tabled data. The downstream classifier is intentionally constrained to use only statistical descriptors computed from extracted pitches, enabling both end-to-end feasibility and interpretability in terms of register and pitch dispersion.

For our experiments, we used the PRIMUS dataset, a large collection of sheet music images widely used in OMR research. Each example in the dataset is identified with one of the three main clefs: *Csol*, *Cdo*, and *Cfa*. The goal is to test whether we can predict which clef should be there using only the statistics of the remaining notes. If this works, we will have a reliable method to "fill in the gap" in cases where the original clef has disappeared or is illegible due to wear and tear on the sheet music.

We processed the sheet music images using the YOLOv11 model, trained to identify the clefs (*Cdo*, *Cfa*, *Csol*) and the position of the notes. We configured the detector with a resolution of 640 pixels, a balance we found to ensure the computer can see small musical symbols without requiring excessive processing. For each item found, the model generates a location frame and a confidence score; to avoid errors and false positives, we only accepted detections where the system was at least 70% certain.

Our approach transforms the position of the notes found into numbers from 0 to 21, which represent the pitch of each note on the staff. This conversion transforms the image into a list of numerical values, retaining only the pitch information necessary to deduce the clef. It is important to note that, at this stage, the term "pitch" refers only to the physical position of the note on the lines and spaces; the actual musical meaning (whether the note is an A or a C) will only be defined after we discover the correct clef.

For each section of the score, we take the positions of the notes and calculate 10 statistical measures that summarize how they are distributed. These characteristics (listed in Table 1) serve as a summary of the notes' behavior: we include the extremes (*minimum* and *maximum*), the center of the distribution (*mean*

and *median*), how much they vary (*standard deviation* and *range*), as well as indicators of note volume and asymmetry. We use standard mathematical definitions to ensure that the calculations are accurate and easy to reproduce.

The resulting dataset contains 950 labeled samples and exhibits an imbalanced class distribution typical of real-world scores: *Csol* (67.7%), *Cdo* (28.6%), and *Cfa* (3.7%).

Table 1. Calculated statistical meta-features from pitch indices.

#	Feature	Description
1	maximum	Maximum pitch index in the segment
2	minimum	Minimum pitch index in the segment
3	mean	Arithmetic mean of pitch indices
4	median	Median of pitch indices
5	standard_deviation	Dispersion of the pitch distribution
6	pitch_count	Number of detected pitches (n)
7	range	Difference between maximum and minimum
8	q1	First quartile (25%)
9	q3	Third quartile (75%)
10	skewness	Asymmetry of the pitch distribution

3.1 Preprocessing and Train–Test Partitioning

Before model training, feature vectors are standardized using StandardScaler. For each feature dimension x, scaling is computed as

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ are estimated on the training split only, preventing information leakage. Data is partitioned into 80% training and 20% test sets using stratification to preserve the empirical class proportions.

Three supervised learning models commonly used for tabular classification are compared in k-nearest neighbors, random forest and gradient boosting. Hyperparameters are selected via grid search on the training data. Specifically, k-nearest neighbors explores neighborhood sizes $k \in \{3, \ldots, 15\}$ and alternative distance metrics; random forest varies the number of trees, maximum depth, and split criteria; and gradient boosting evaluates combinations of learning rate, tree depth, and number of estimators. This procedure ensures that each model is evaluated under a competitive configuration rather than an arbitrary default.

We performed the model selection using stratified 5-fold cross validation with shuffling enabled and a fixed random state (42) to ensure reproducibility. Because

the dataset is markedly imbalanced, we adopt the weighted F1-score as the primary evaluation metric. We denoted $F1_c$ as the class-wise F1-score and w_c as the support (relative frequency) of class c; the weighted F1 we computed as

$$\sum_c w_c \, F1_c,$$

providing an effective assessment of performance, but without overestimating the accuracy of the majority class. For greater comprehensiveness, we also report accuracy on the test set and training time in the results section to characterize the relationship between accuracy and efficiency among the methods.

4 Discussion of Results

According to our analysis, the grid search results revealed a superiority of gradient boosting compared to the other algorithms tested. As shown in Table 2, gradient boosting achieved a weighted F1-score of 0.7242 on the test set, surpassing k-nearest neighbors (0.6950) and random forest (0.6898). This improvement is also reflected in the higher test accuracy (0.7316), although with a higher training cost (61.24 s). Overall, we believe these results indicate that additive sets of shallow decision trees provide an effective inductive bias for key inference under height distribution meta-features, capturing non-linear decision boundaries induced by quartiles, dispersion, and central tendency statistics.

Given the grid search results, the best configuration we found for gradient boosting was a learning rate of 0.1, a maximum depth of 3, a minimum sample split of 2, 50 estimators, and full subsampling (1.0), suggesting a conservative capability that balances expressiveness of learning and prevention of overfitting. This is consistent with the task structure: although clef selection is strongly linked to musical register, its discriminative cues rarely reduce to a single threshold of pitch values and instead rely on distributional signatures that benefit from non-linear combinations. Figure 4 complements Table 2 illustrating the comparative behavior of the models with respect to the reported metrics.

Table 2. Comparative results of machine learning models for clef prediction.

Model	F1-Score (CV)	F1-Score (Test)	Accuracy (Test)	Time (s)
Gradient Boosting	0.6802	**0.7242**	**0.7316**	61.24
KNN	0.6954	0.6950	0.7105	3.05
Random Forest	0.6886	0.6898	0.7105	40.54

In addition to being efficient, this model allows us to understand exactly what it takes into account to decide which clef it is. By analyzing the importance of the features (Table 3), we note that the third quartile (q_3) is the most influential

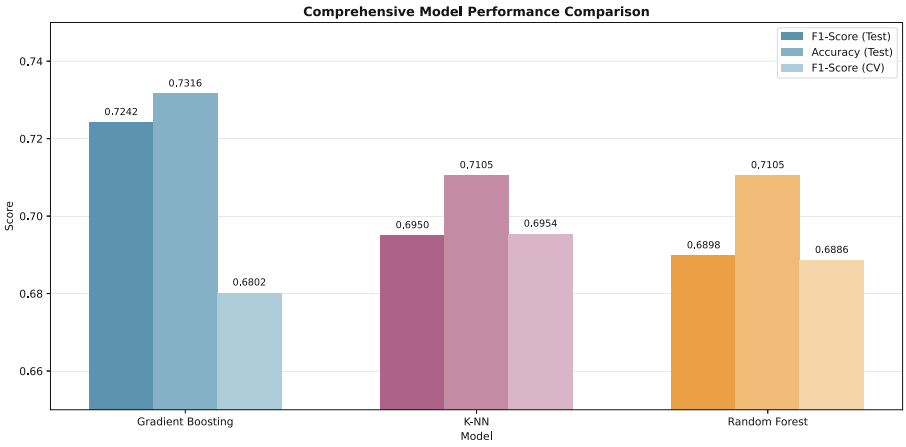


Fig. 4. Comparison of performance metrics in clef prediction.

factor, followed by the standard deviation and the median. This shows us that, to identify the clef, the model prefers to observe where the highest part of the notes is concentrated and how much they vary, rather than focusing only on extreme values or an isolated average.

The maximum and minimum values of the notes had almost no importance (0.0093 and 0.0003), which proves that a single very high or low note is not sufficient to reliably predict the clef. Because of this, we removed these two characteristics before the final training, simplifying the model without losing quality. This result makes perfect sense from a musical point of view: clefs serve to indicate the region where the music "lives" (the register and the tessitura), something that becomes much clearer through the concentration of the highest

Table 3. Feature importance in the Gradient Boosting model.

Rank	Feature	Importance
1	q3	0.2940
2	standard_deviation	0.1837
3	median	0.1288
4	mean	0.1222
5	skewness	0.1097
6	q1	0.0908
7	pitch_count	0.0368
8	range	0.0244
9	maximum	0.0093
10	minimum	0.0003

notes (q_3) and their variation than through extreme notes that appear only occasionally.

5 Conclusion and Future Work

In this article, we tested whether it was possible to recover musical clefs by combining computer vision and data analysis. We created a workflow that first uses YOLOv11 to detect clefs and notes, transforms these images into numbers, and summarizes each section of the score with 10 simple statistical measures. To test whether these numbers were sufficient to predict the clef, we evaluated three algorithms (k-NN, Random Forest, and Gradient Boosting), focusing on the F1 score to ensure fair results, since some clefs appear much more frequently than others in our dataset.

Our analyses showed that Gradient Boosting was the most efficient (F1 of 0.7242), confirming that notes alone already give very strong clues about which is the correct clef. In addition to correctly predicting the result, the model gave us a logical explanation: it prioritizes the region where the highest notes are concentrated and how much they vary, rather than focusing on isolated extreme notes. This makes perfect musical sense, since the clef serves precisely to define the region (or tessitura) in which the music should be read.

For the future, we have three main paths. First, we want to expand the system to identify other essential symbols, such as key signatures and time signatures. Second, we intend to test the model on real historical manuscripts, which are much more challenging than digital images. Finally, we want to test whether including small melodic or rhythmic excerpts can help the model become even more accurate, combining the statistics we already use with the musical context of each passage.

Acknowledgment. The authors acknowledge partial funding from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the National Council for Scientific and Technological Development - Brazil (CNPq) - Funding Codes 442344/2023-6, 444602/2024-0, and 407466/2025-8; the Espírito Santo Research and Innovation Support Foundation (FAPES)–Call No. 02/2026 (Participation in Technical-Scientific Events)–Grant No. 428/2026 and project #2024-9G1WP BPC; and the Minas Gerais Research Funding Foundation (FAPEMIG)

References

1. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), pp. 591–596 (2011). https://ismir2011.ismir.net/papers/OS6-1.pdf
2. Calvo-Zaragoza, J., Rizo, D.: End-to-end neural optical music recognition of monophonic scores. Appl. Sci. **8**(4), 606 (2018)

3. Chew, E.: Mathematical and Computational Modeling of Tonality: Theory and Applications. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-9475-1
4. Huang, C.Z.A., et al.: Music transformer: generating music with long-term structure. In: International Conference on Learning Representations (ICLR) (2019). https://openreview.net/forum?id=rJe4ShAcF7
5. Pacha, A., Jr., J.H., Calvo-Zaragoza, J.: A baseline for general music object detection with deep learning. Appl. Sci. **8**(9), 1488 (2018). https://doi.org/10.3390/app8091488
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). https://openaccess.thecvf.com/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html
7. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7464–7475 (2023). https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html

A Multivariate Multi-step Deep Learning Framework for InSAR Displacement Prediction

Maria Maia[1(✉)], Luís Miguel Matos[1,2], Luís Magalhães[1,3], Joaquim Tinoco[4], and Steffan Davies[5]

[1] ALGORITMI/LASI Research Center – Department Information Systems, Minho University, Guimarães, Portugal
pg55904@alunos.uminho.pt, {luis.matos,lmagalhaes}@dsi.uminho.pt
[2] EPMQ, CCG ZGDV Institute, Guimarães, Portugal
[3] CVIG, CCG ZGDV Institute, Guimarães, Portugal
[4] ISISE, ARISE, Department of Civil Engineering, University of Minho, Guimarães, Portugal
jtinoco@civil.uminho.pt
[5] Spotlite-AETHRA, Rua Pedro Nunes, IPN, C 3030-199 Coimbra, Portugal
steffan@spotlitedata.com

Abstract. Accurate forecasting of slope displacements is critical for landslide risk management and the development of reliable early warning systems. This study investigates the application of three deep learning architectures: Temporal Convolutional Network (TCN), Patch Time Series Transformer (PatchTST), and Temporal Fusion Transformer (TFT) for multi-step forecasting of InSAR-derived slope displacements across multiple prediction horizons. The proposed framework adopts a joint multivariate representation that simultaneously predicts displacements across all monitored points, incorporating spatial context through neighbouring measurements and daily precipitation as an environmental covariate. Models are evaluated on both Persistent Scatterer (PS) and Small Baseline Subset (SBAS) datasets, considering horizontal and vertical displacement components independently. Results show that all three models achieve R^2 values consistently above 0.75, with TFT demonstrating the strongest overall performance on PS data and PatchTST reaching the highest R^2 of 0.91 on the SBAS vertical component at $H = 5$. The TCN, while generally less competitive at longer horizons, outperforms attention-based models on the SBAS horizontal component, suggesting that its convolutional inductive bias is better suited to signals exhibiting rapid directional reversals. These findings highlight the complementary strengths of the investigated architectures and establish a foundation for future integration of additional environmental covariates and more rigorous validation strategies.

Keywords: Slope Displacement Prediction · InSAR · Deep Learning · Network Level Scalability

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 639–650, 2027.
https://doi.org/10.1007/978-3-032-30494-0_40

1 Introduction

Slope instabilities, including landslides, rock falls, and debris flows, pose significant hazards in both natural and engineered slopes across all geographic regions of the world. These movements can be triggered by natural factors such as intense rainfall, rapid snowmelt, earthquakes, and volcanic eruptions, as well as by human activities such as excavation and road construction [1]. The resulting hazards are responsible for thousands of deaths and injuries, as well as monetary losses [2], highlighting the need for effective investigation, monitoring, and remediation measures.

Advanced remote sensing techniques, including Interferometric Synthetic Aperture Radar (InSAR) together with Machine Learning (ML) approaches, have significantly improved the detection, monitoring and forecasting of slope displacements [3,4]. InSAR techniques such as PS-InSAR [5] and SBAS-InSAR [6] enable systematic, repeated observations of ground deformation over large areas with millimetric precision, supporting both rapid post-event mapping and long-term monitoring of slow-moving slopes [7].

Building upon these techniques, ML, and more specifically Deep Learning (DL), have emerged as powerful tools to further enhance displacement forecasting. Given that InSAR derived measurements are inherently acquired as multitemporal time series, DL architectures designed for sequential data have shown strong capability in capturing the complex and nonlinear temporal patterns of slope displacements. The combination of these approaches has therefore opened new perspectives for the development of more accurate and reliable early warning systems, contributing to improved landslide risk management at both local and regional scales [8].

Motivated by the limited exploration of advanced DL architectures in the geotechnical domain, this study investigates three models with distinct and complementary strengths: the Temporal Fusion Transformer (TFT) [9], adopted for its ability to handle heterogeneous multivariate inputs while providing interpretable insights through its attention mechanism; the Patch Time Series Transformer (PatchTST) [10], selected for its patch-based tokenisation that captures local temporal semantics and attends to longer historical windows with reduced computational cost; and the Temporal Convolutional Network (TCN) [11], included as a strong convolutional baseline, leveraging dilated causal convolutions for efficient modelling of long-range temporal dependencies. All three models are evaluated under a unified multi-step forecasting framework across multiple prediction horizons, considering all monitored points simultaneously, where the spatial context of neighbouring points at a given temporal *lag* is incorporated as input features.

The remainder of this paper is organized as follows: Sect. 2 presents a review of the related work on DL approaches for landslide displacement forecasting, Sect. 3 describes the proposed methodology, Sect. 4 presents the experimental results and Sect. 5 draws the main conclusions, while outlining directions for future work.

2 Related Work

The application of DL to the forecasting of displacement derived from InSAR data has grown substantially over the past decade.

Building on this foundation, Mirmazloumi et al. [12] coupled point-scale LSTM forecasting with a spatial early alert detector that considers adjacent measurement points, demonstrating that spatial context reduces false alarm rates in mining collapse scenarios. Wang et al. [13] proposed the Spacetimeformer network, which jointly attends to temporal and spatial relationships across InSAR points via a dual global-local self-attention mechanism, outperforming LSTM and vanilla Transformer baselines, though at the cost of extensive site-specific preprocessing that constrains scalability. Moualla et al. [14] addressed irregular Sentinel-1 temporal sampling, showing that TFT achieves the best trade-off between accuracy and computational cost, but treating each monitoring point independently without exploiting spatial structure.

Taken together, these studies reveal a clear tension: approaches that model spatial dependencies tend to require regular temporal sampling and site-specific preprocessing, while methods designed for irregular time series treat each point independently. Additionally, spatial context is often incorporated heuristically through neighbourhood averaging [12], rather than through learned spatiotemporal representations.

The present study addresses this gap by investigating three architecturally distinct deep learning models under a unified multi-step forecasting framework that simultaneously predicts displacements across all monitored points and incorporates spatial context through neighbour measurements at a given temporal lag as explicit input features. This work extends the preliminary investigation of Owusu-Ansah et al. [15], who applied MR and ANN to InSAR vertical displacement time series from highway slopes using a sliding-window strategy, achieving $R^2 = 0.93$ but identifying the need for additional input variables and more robust model optimization.

3 Materials and Methods

3.1 Data

InSAR Displacement Data. The data used in this study were derived from Sentinel-1 acquisitions processed via PS-InSAR and SBAS-InSAR techniques. The resulting datasets are organized into two resolution categories: Persistent Scatterer (PS) and Small Baseline Subset (SBAS), each containing separate files for horizontal and vertical displacement components, further stratified by monitored element: slope, pavement and unclassified area of interest (AOI). Each file comprises 88 columns, including longitude, latitude and the displacement time series spanning from 5 June 2023 to 16 November 2025. The number of measurement points per file varies with both resolution and element type, as summarised in Table 1.

Table 1. Dimensions of the IP3 InSAR datasets.

Category	PS rows	SBAS rows
Slope	38	34
Pavement	92	89
AOI	933	671

Precipitation Data. Precipitation data were obtained from the PDIR-Now product, available through the CHRS Data Portal [16]. Daily accumulated precipitation values (mm) were retrieved in NetCDF format for the period coinciding with the InSAR time series. Since the PDIR-Now grid resolution does not align directly with the InSAR measurement point coordinates, a nearest-neighbour spatial matching procedure was applied to associate each point with its closest grid cell. The resulting data were then restructured into a tabular format and subsequently merged with the corresponding InSAR displacement files on the basis of matching spatial coordinates, yielding a unified dataset ready for model input.

3.2 Methods

All models share a common input representation and training pipeline. The displacement time series from all monitored points are first reorganised into a sliding-window format, extracting sequences of $K = 4$ consecutive observations as input and the following H observations as prediction targets, a configuration empirically determined in [15] as the best-performing baseline. This windowing is applied independently to each point and each forecasting horizon $H \in \{1, 5, 10\}$, corresponding to prediction windows of approximately 12, 60 and 120 days, respectively. Rather than treating each monitoring point independently, the framework constructs a joint multivariate representation capturing the spatial context of the entire network, with neighbouring points selected based on geographic proximity to the point of interest and modelled through direct correlations without distance-based weighting. At each time step K, the displacement values of all monitored points are stacked into a single vector and the corresponding precipitation values are concatenated alongside them, producing a combined feature vector of dimension $n_{\text{points}} \times 2$ per time step. This structure is repeated across the $K = 4$ lag steps, resulting in 3D input arrays of shape $(\ n_{\text{samples}},\ 4,\ n_{\text{points}} \times 2)$. The target arrays are constructed analogously, stacking the H future displacement values for all points into a flat output vector of dimension $H \times n_{\text{points}}$. Model validation is assessed through 5-fold cross-validation without shuffling, preserving the temporal ordering of the data, and the best-performing fold is retrained on the full dataset to produce final predictions.

The three main architectures investigated are the TCN, TFT, and PatchTST. The TCN [11] replaces recurrence with stacked dilated causal convolutions,

enabling the capture of long-range temporal dependencies while maintaining computational efficiency and avoiding the vanishing gradient problems of recurrent models. The architecture comprises a single TCN block with $F = 64$ filters, a kernel size of 3, and four dilation levels $\{1, 2, 4, 8\}$, with causal padding and skip connections. A dropout rate of 0.1 is applied within the block, and the final representation is passed to a dense output layer with linear activation. Training is conducted with the Adam optimiser and MSE loss, with early stopping over a maximum of 100 epochs and a batch size of 32. The TFT [9] combines a multi-head self-attention mechanism with position-wise feed-forward layers and residual connections, offering interpretable multi-horizon forecasting through its attention weights. The input sequence is first projected to a latent dimension of $d_{\text{model}} = 64$, followed by a multi-head attention layer with $n_{\text{heads}} = 4$ and a feed-forward network of dimension 128 with ReLU activation. Residual connections and layer normalisation are applied after both sub-layers, and a global average pooling operation aggregates the temporal representations before a dropout layer and the final output projection. The PatchTST [10] extends this Transformer encoder by introducing a patch-based tokenisation of the input sequence prior to the attention mechanism: a 1D convolutional layer with patch length 2 and $F = 64$ filters segments the input into overlapping sub-series patches reducing the effective sequence length while preserving local temporal structure. This design is particularly advantageous for longer horizons, as the patching step reduces the quadratic cost of self-attention. Both Transformer-based models are trained under the same protocol as the TCN.

The predictive performance of all models is assessed using three metrics: the coefficient of determination (R^2), Mean Absolute Error (MAE) and Normalised MAE (NMAE), defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{NMAE} = \frac{\text{MAE}}{\max(y) - \min(y)} \tag{3}$$

where y_i denotes the observed displacement value, $\hat{y}_i$ the corresponding predicted value, $\bar{y}$ the mean of observed values, and n the total number of samples.

4 Results

Figure 1 presents the time series comparison between observed and predicted vertical displacements for a representative PS monitoring point for $H = 1$[1].

[1] Time series comparisons (Figs. 1, 3, 5 and 7) are shown for $H = 1$ only, as qualitative behaviour is consistent across all horizons.

All three models closely track the observed signal, capturing both the gradual negative trend and the more abrupt displacement events characteristic of the mid-2024 period. The TFT and PatchTST predictions exhibit smoother trajectories compared to the TCN, which tends to follow higher-frequency fluctuations more closely, though at the cost of occasional over-responses to short-term noise.

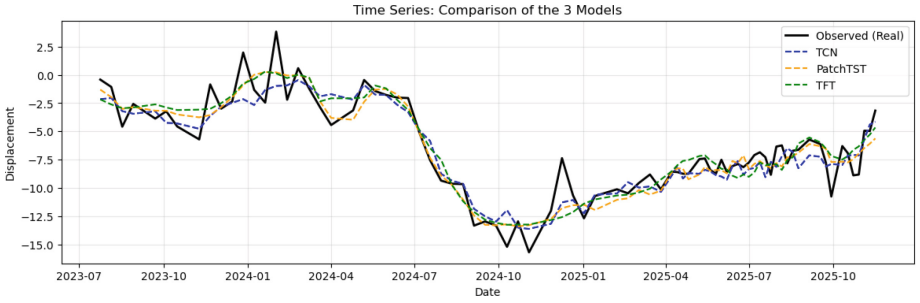


Fig. 1. Time series comparison of observed and predicted PS vertical displacements for $H = 1$.

The quantitative evaluation across all three horizons is summarised in Fig. 2. The TFT consistently achieves the lowest MAE and NMAE values across all forecasting horizons, while also attaining the highest R^2 scores, particularly at $H = 5$ and $H = 10$ where it reaches approximately 0.88 and 0.78, respectively. PatchTST performs comparably to TFT at $H = 1$ but shows a slight degradation at longer horizons. The TCN, while competitive at $H = 1$ ($R^2 \approx 0.85$), exhibits a more pronounced decline in performance as the horizon extends, suggesting reduced capacity to capture multi-step displacement dynamics relative to the attention-based architectures.

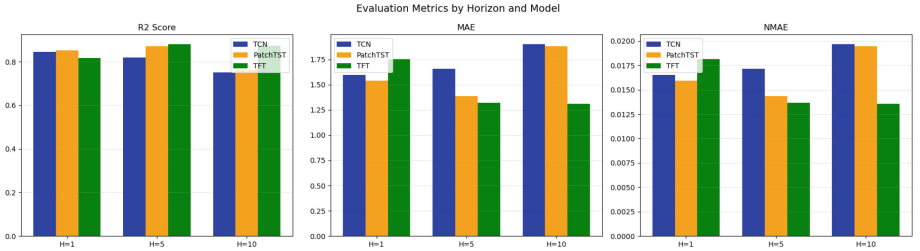


Fig. 2. Quantitative evaluation across all horizons for PS vertical displacements.

The time series comparison for PS horizontal displacement for $H = 1$, shown in Fig. 3, reveals a broadly similar pattern to the vertical component, with a pronounced negative trend developing from early 2024 and partial recovery towards

the end of the monitoring period. However, the horizontal signal exhibits greater high-frequency variability and more abrupt short-term fluctuations, posing a more challenging forecasting target. The TFT and PatchTST models produce smoother predictions that capture the dominant trend effectively, while the TCN tends to lag slightly behind rapid directional changes, particularly during the transition period between mid-2023 and early 2024.

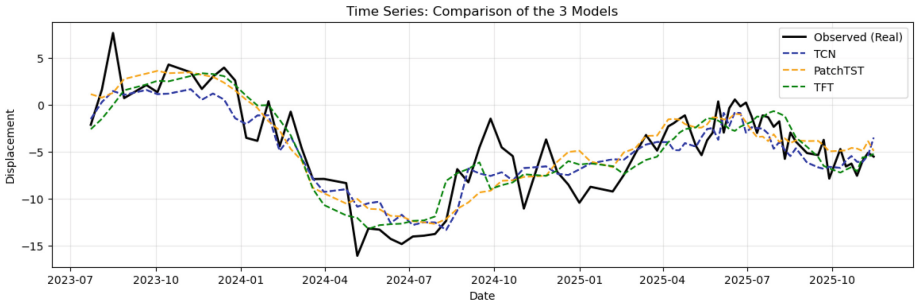


Fig. 3. Time series comparison of observed and predicted PS horizontal displacements for $H = 1$.

The evaluation metrics presented in Fig. 4 confirm that the horizontal component is generally harder to predict than the vertical, as reflected by slightly higher MAE and NMAE values across all models and horizons. Nonetheless, the ranking of models remains consistent: TFT achieves the lowest error and highest R^2 across all horizons, followed closely by PatchTST, with the TCN exhibiting higher MAE values. A notable observation is that the TCN's R^2 score remains relatively stable across horizons (≈ 0.85), suggesting that while its absolute error is higher, its ability to explain variance in the horizontal displacement does not degrade as sharply with increasing horizon as observed in the vertical component.

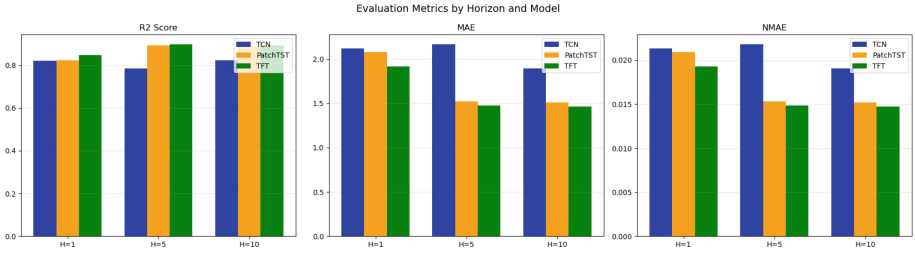


Fig. 4. Quantitative evaluation across all horizons for PS horizontal displacements.

The SBAS vertical displacement time series for $H = 1$, shown in Fig. 5, reveals a more pronounced and sustained negative trend than its PS counterpart, with a sharp acceleration in displacement occurring around mid-to-late

2024, reaching values below −25 mm before a partial recovery in early 2025. This abrupt event represents the most challenging feature of the entire dataset, and the models' ability to anticipate it varies considerably. The PatchTST and TFT predictions track the gradual pre-event trend well but underestimate the magnitude of the rapid displacement episode, smoothing over the most extreme values. The TCN, by contrast, exhibits a more erratic response during this period, with predictions diverging more significantly from the observed signal, suggesting reduced robustness to sudden regime changes.

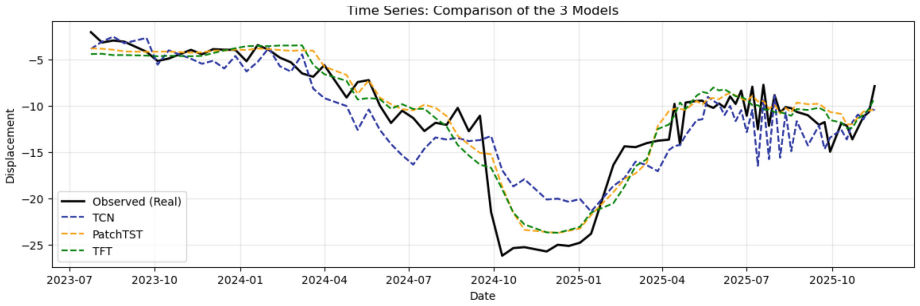


Fig. 5. Time series comparison of observed and predicted SBAS vertical displacements for $H = 1$.

The quantitative metrics in Fig. 6 reflect this increased difficulty. While R^2 values remain broadly comparable to the PS results ($\approx 0.80 - 0.90$ across models and horizons), the absolute MAE values are higher, with the TCN reaching ≈ 2.3 at $H = 1$ compared to ≈ 1.8 for TFT. A particularly noteworthy result is that PatchTST achieves its best relative performance at H = 5, surpassing TFT in R^2 score, suggesting that its patch-based tokenisation may be especially advantageous for capturing medium-range dynamics. At $H = 10$, all models show some degradation, though TFT and PatchTST maintain lower error levels than the TCN.

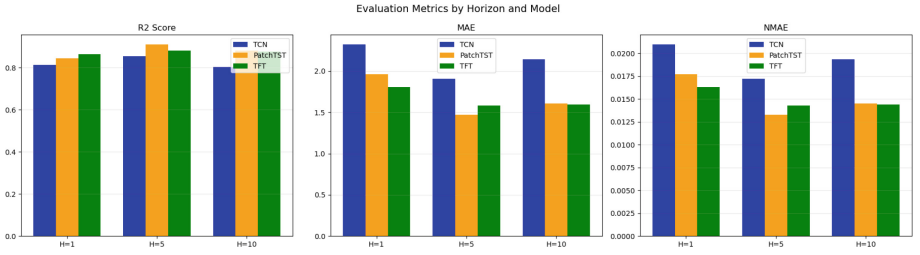


Fig. 6. Quantitative evaluation across all horizons for SBAS vertical displacements.

The SBAS horizontal displacement for $H = 1$, shown in Fig. 7, presents a distinct behavioural pattern relative to the other components, characterised by an initial phase of moderate positive displacements followed by a sharp negative excursion around mid-to-late 2024 and a strong positive recovery trend through 2025. This signal exhibits the highest degree of non-stationarity across all four datasets, with both the direction and rate of displacement reversing markedly within the monitoring period. Despite this complexity, TFT and PatchTST broadly capture the dominant trend transitions, though they tend to underestimate peak values during the most dynamic phases. The TCN, while tracking the general trajectory, shows greater variability in its predictions and occasional phase misalignment during abrupt reversals.

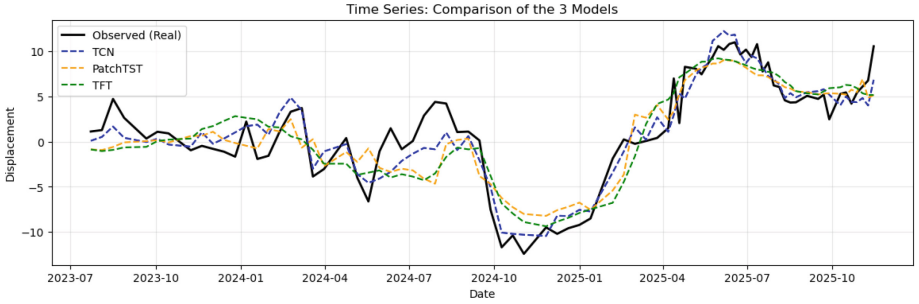


Fig. 7. Time series comparison of observed and predicted SBAS horizontal displacements for $H = 1$.

The evaluation metrics presented in Fig. 8 highlight an important reversal in model rankings for this component: the TCN achieves the highest R^2 scores at all horizons (≈ 0.90 at $H = 1$ and $H = 10$), and the lowest MAE at $H = 1$ (≈ 1.58), while PatchTST and TFT exhibit notably higher errors at the shortest horizon, with MAE values reaching ≈ 2.08. This suggests that the convolutional inductive bias of the TCN may be better suited to capturing the more oscillatory and rapidly reversing dynamics of the SBAS horizontal signal, whereas the attention-based models.

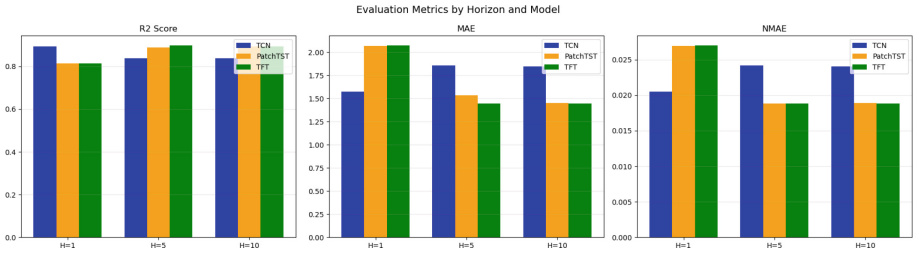


Fig. 8. Quantitative evaluation across all horizons for SBAS horizontal displacements.

5 Conclusions

This study investigated the application of three architectures: TCN, PatchTST and TFT for multi-step forecasting of InSAR-derived slope displacements across PS and SBAS datasets for both vertical and horizontal components. All three models demonstrated reasonable predictive capability, achieving R^2 values consistently above 0.75 across all configurations, with TFT and PatchTST generally outperforming the TCN at medium and long forecasting horizons. The TFT exhibited the strongest overall performance on the PS datasets, while PatchTST achieved the highest R^2 of all models across all configurations on the SBAS vertical component at $H = 5$ (0.91). The TCN, despite being generally less competitive, proved to be the best-performing model on the SBAS horizontal component at $H = 1$, suggesting that its convolutional inductive bias may be advantageous for signals characterised by rapid directional reversals, where the smoothing tendency of attention-based models becomes a limitation. The joint multivariate representation across all monitored points, combined with the inclusion of precipitation as an environmental covariate, represents a methodological contribution relative to purely univariate approaches. However, some limitations should be acknowledged, including the short input window of $K = 4$ that may constrain the models' ability to capture longer-period seasonal or climatic cycles that influence displacement dynamics observations and the preliminary nature of the hyperparameter optimisation. Future work will focus on the integration of additional environmental covariates and on the adoption of more rigorous model validation strategies, along with more exhaustive hyperparameter optimisation and the incorporation of explainability techniques to better assess the generalisation capacity and interpretability of the proposed frameworks.

Acknowledgments. This work was funded by the ANI (Agência Nacional de Inovação) through the R&D project "WISE - Wireless Transportation Infrastructure Safety Evaluation", project number 14383, operation code COMPETE2030-FEDER-00574700, cofunded by the Innovation and Digital Transition Program (COMPETE 2030), by the PORTUGAL 2030 and by the European Union. Also, this work is financed by national funds through FCT–Foundation for Science and Technology, under grant agreement CEECINST/00018/2021/CP2806/CT0006 attributed to the 2nd author. This work was also supported by FCT/MCTES under the R&D Unit Institute for Sustainability and Innovation in Structural Engineering (ISISE), under the references UID/4029/2025 (https://doi.org/10.54499/UID/04029/2025), UID/PRR/04029/2025 (https://doi.org/10.54499/UID/PRR/04029/2025) and UID/PRR2/04029/2025 (https://doi.org/10.54499/UID/PRR2/04029/2025), and under the Associate Laboratory Advanced Production and Intelligent Systems ARISE under reference LA/P/0112/2020 (https://doi.org/10.54499/LA/P/0112/2020).

References

1. Schuster, R.L.: Socioeconomic significance of landslides. In: Turner, A.K., Schuster, R.L. (eds.) Landslides - Investigation and Mitigation. Transportation Research Board Special Report 247, pp. 12–35. National Academy Press, Washington, DC (1996)
2. Petley, D.N.: Landslides and engineered slopes: protecting society through improved understanding. In: Proceedings of the 11th International Symposium on Landslides and 2nd North American Symposium on Landslides, pp. 3–13. CRC Press, Taylor and Francis Group, Boca Raton (2012)
3. Casagli, N., Cigna, F., Bianchini, S., et al.: Landslide mapping and monitoring by using radar and optical remote sensing: examples from the EC-FP7 project SAFER. Remote Sens. Appl. Soc. Envir. **4**, 92–108 (2016). https://doi.org/10.1016/j.rsase.2016.07.001
4. Nava, L., Carraro, E., Reyes-Carmona, C., et al.: Landslide displacement forecasting using deep learning and monitoring data across selected sites. Landslides **20**(10), 2111–2129 (2023). https://doi.org/10.1007/s10346-023-02104-9
5. Ferretti, A., Prati, C., Rocca, F.: Permanent scatterers in SAR interferometry. IEEE Trans. Geosci. Remote Sens. **39**(1), 8–20 (2001)
6. Berardino, P., Fornaro, G., Lanari, R., Sansosti, E.: A new algorithm for surface deformation monitoring based on small baseline differential SAR interferometry. IEEE Trans. Geosci. Remote Sens. **40**(11), 2375–2383 (2002)
7. Novellino, A., Pennington, C., Leeming, K., et al.: Mapping landslides from space: a review. Landslides **21**(5), 1041–1052 (2024). https://doi.org/10.1007/s10346-024-02215-x
8. Ebrahim, K.M.P., Fares, A., Faris, N., Zayed, T.: Exploring time series models for landslide prediction: a literature review. Geoenvironmental Disasters **11**(25), 1–28 (2024). https://doi.org/10.1186/s40677-024-00288-3
9. Lim, B., Arık, S.Ö., Loeff, N., Pfister, T.: Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. Int. J. Forecast. **37**(4), 1748–1764 (2021). https://doi.org/10.1016/j.ijforecast.2021.03.012
10. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In: 11th International Conference on Learning Representations, ICLR 2023 (2023). https://doi.org/10.48550/arXiv.2211.14730
11. Bai, s., kolter, j.z., koltun, v.: an empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv preprint arXiv:1803.01271 (2018). https://doi.org/10.48550/arXiv.1803.01271
12. Mirmazloumi, S.M., Wassie, Y., Nava, L., Cuevas-González, M., Crosetto, M., Monserrat, O.: InSAR time series and LSTM model to support early warning detection tools of ground instabilities: mining site case studies. Bull. Eng. Geol. Environ. **82**, 374 (2023). https://doi.org/10.1007/s10064-023-03388-w
13. Wang, J., et al.: Spatiotemporal mechanism-based Spacetimeformer network for InSAR deformation prediction and identification of retrogressive thaw slumps in the Chumar River Basin. Remote Sens. **16**(11), 1891 (2024). https://doi.org/10.3390/rs16111891
14. Moualla, L., Rucci, A., Naletto, G., Anantrasirichai, N., Da Deppo, V.: Hybrid GIS-transformer approach for forecasting Sentinel-1 displacement time series. Remote Sens. **17**(14), 2382 (2025). https://doi.org/10.3390/rs17142382

15. Owusu-Ansah, D., Tinoco, J., Matos, J., Cabral, R.: Combining slope satellite image analysis and artificial intelligence algorithms for highway service level assessment. In: Proceeding XVIII ECSMGE 2024, pp. 1269–1272. Taylor & Francis, London (2024). https://doi.org/10.1201/9781003431749-230
16. CHRS Data Portal, University of California, Irvine. https://chrsdata.eng.uci.edu/. Accessed 16/Mar/2026

Computational Assessment of Greenness: Objective vs. Perceived Greenness in Urban Green Spaces

Amirmohammad Ghavimi(✉) and Roland Pesch

Jade University of Applied Sciences, 26121 Oldenburg, Germany
amirmohammad.ghavimi@jade-hs.de

Abstract. Urban public green spaces (UPGSs) are key elements of sustainable cities, yet "greenness" can be assessed both through objective remote-sensing indicators and through residents' perceptions, which do not necessarily align. This pilot study compares satellite-derived greenness with perceived greenness for UPGSs in Oldenburg (Lower Saxony, Germany). Objective greenness was quantified using mean NDVI computed from Sentinel-2 imagery (B04/B08) for 2024–2025, aggregated monthly within ATKIS-derived UPGS polygons while excluding buildings and water bodies. Perceived greenness was obtained via an online survey (n = 109) using a Likert scale (0–7); UPGSs with at least two ratings were retained (n = 24). Agreement between objective and perceived greenness was assessed using rank-based correlation. Results indicate a weak positive but non-significant association between mean NDVI and perceived greenness (*Kendall's* τ_b=0.22304, p = 0.130; *Spearman's* ρ=0.276, p = 0.191), suggesting limited alignment in this pilot dataset. The study provides a transparent computational workflow and baseline evidence to support future perception-aware urban greening assessment with larger samples and additional UPGS attributes.

Keywords: Urban public green spaces · Greenness · NDVI · Perceived greenness · Remote sensing · Urban sustainability

1 Introduction

UPGS are increasingly framed as multifunctional urban infrastructure that supports environmental performance, social cohesion, and public health. Beyond their ecological roles (e.g., microclimate regulation, stormwater retention, biodiversity support), UPGSs contribute to livability and long-term urban resilience—core concerns of sustainable development agendas and urban policy [1–3]. In sustainability-oriented planning, "greenness" is often used as a headline indicator because it can summarize vegetation presence while also relating to how green spaces are experienced by residents. Greener UPGSs are frequently associated with higher usage, improved mental well-being, and beneficial social outcomes, and broader evidence links green space exposure and availability to reduced mortality risk [4, 5]. These connections align with the Sustainable Development Goals (SDGs), particularly SDG 11 (Sustainable Cities and Communities) and SDG 3 (Good

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 651–661, 2027.
https://doi.org/10.1007/978-3-032-30494-0_41

Health and Well-Being), and motivate planning tools that can evaluate urban greenness in ways that are both scalable and socially meaningful.

From a planning and governance perspective, perceived greenness is not merely a subjective "opinion" but a planning-relevant indicator because it is closely related to residents' satisfaction, place attachment, perceived restorative potential, and willingness to use and support green infrastructure interventions. Perceptions reflect how people interpret the quality and character of green environments (e.g., naturalness, aesthetics, cleanliness, safety, and facilities), which can mediate whether UPGSs deliver expected social and health benefits [6, 7]. Accordingly, perceived greenness can be viewed as a component of social sustainability: if certain areas systematically experience lower perceived greenness—despite nominal vegetation presence—then policies based only on biophysical indicators may overlook inequalities in lived experience and, potentially, inequalities in health and well-being outcomes. Recent work highlights those distributional measures of green space alone are insufficient unless paired with indicators closer to human experience and use [6].

At the same time, cities need operational, scalable metrics to monitor greenness of UPGSs over large areas and over time. Remote sensing provides such a computational basis. The Normalized Difference Vegetation Index (NDVI) is one of the most widely used indicators for mapping greenness, because it can be derived consistently from satellite imagery and is sensitive to vegetation presence and condition [8, 9]. NDVI has therefore become common in planning support systems and research on urban greening, exposure assessments, and policy evaluation. However, NDVI primarily captures spectral properties of vegetation from a "top-down" perspective and may not reflect what people visually perceive or value on the ground. Empirical studies repeatedly report only moderate or variable agreement between NDVI-based greenness and self-reported or perceived greenness [10–12]. This divergence is expected because objective and perceived greenness represent different constructs: perception can depend on visibility and structure of vegetation (trees vs. grass), spatial configuration, maintenance, and contextual cues, while objective indices can be sensitive to scale, seasonality, and land-cover composition [8, 13]. To bridge this gap, recent computational work has explored complementary indicators such as street-level "green view" measures or multi-source assessment pipelines that better approximate human-scale exposure [13–15]. Nevertheless, NDVI remains attractive for planning due to its accessibility and repeatability, making it important to quantify when NDVI provides a reasonable proxy for perceived greenness and when it does not.

While mismatches between satellite-derived greenness and perceived greenness have been reported in prior work, evidence remains context-dependent and highly sensitive to how green spaces are delineated and how greenness is aggregated. This study contributes (i) an automated and transparently documented NDVI-processing workflow for ATKIS-derived UPGS polygons, including intra-polygon exclusion of water bodies and buildings and monthly aggregation across 2024–2025, and (ii) a UPGS-level comparison in the Oldenburg region that produces directly interpretable rankings and maps for planning support. As a pilot, the paper establishes baseline alignment between objective and perceived greenness and identifies candidate UPGSs for subsequent, factor-based analysis.

This short paper contributes a pilot computational comparison of objective and perceived greenness for UPGSs in the Oldenburg (Germany) region. We compute mean NDVI values (Sentinel 2) within UPGS polygons (excluding water bodies and buildings) and compare the resulting UPGS-level greenness ranking with perceived greenness ratings obtained through an online survey. Using a subset of UPGSs with sufficient survey responses, we quantify the association between objective and perceived greenness via rank-based correlation and provide spatial visualizations of both rankings. As a pilot, the study primarily aims to establish a transparent and well-documented computational workflow and to provide an initial estimate of alignment between satellite-derived greenness and resident perception, supporting future perception-aware urban greening and sustainability-oriented decision-making.

2 Methodology

2.1 Study Area and UPGS Definition

The study focuses on the city region of Oldenburg, Germany. To capture UPGSs potentially used by residents beyond the administrative boundary, we considered the administrative extent of Oldenburg with an additional 5 km buffer.

UPGS geometries and land-cover classification were derived from ATKIS (Amtliches Topographisch-Kartographisches Informationssystem), provided by LGLN (Landesamt für Geoinformation und Landesvermessung Niedersachsen). In the current pilot phase, all ATKIS green-space classes within the study area were treated uniformly as UPGSs; differentiation by green-space type (e.g., park-like areas vs. semi-natural areas) are planned for future work.

2.2 Objective Greenness: NDVI Computation from Sentinel-2

Data Source. We used Copernicus Sentinel-2 multispectral imagery accessed via the Copernicus Data Space Ecosystem. Sentinel-2 provides spatial resolution suitable for urban vegetation analyses, with 10 m pixels for the red (B04) and near-infrared (B08) bands used to compute NDVI.

Image Selection and Pre-processing. Sentinel 2 multispectral imagery (Copernicus Data Space Ecosystem) was analysed for the period 2024–2025. For each month, we selected acquisitions with $< 10\%$ cloud cover (scene-level metadata) for the study area and generated NDVI GeoTIFF outputs within the specified area of interest. NDVI was computed using the standard formulation:

$$NDVI = \frac{NIR - RED}{NIR + RED} = \frac{\mathrm{B08} - \mathrm{B04}}{\mathrm{B08} + \mathrm{B04}} \tag{1}$$

where B04 and B08 are the Sentinel 2 red and near infrared bands, respectively.

To reduce contamination by non-vegetated land covers, water bodies and buildings were excluded from the UPGS polygons (implemented as holes), and NDVI statistics were calculated only over the remaining UPGS area.

While acknowledging potential Areal Unit Problem (MAUP) and ecological fallacy, splitting UPGSs into sub-polygons (e.g., paths) was not undertaken. This is because the study aims to capture the overall perceived greenness, influenced by factors beyond immediate path views (e.g., overall size, surrounding context). Many UPGSs are small, making path-based analysis less meaningful. Restricting analysis to paths also limits the scope, potentially missing the broader influence of the UPGS on perception. Treating the UPGS as a single unit reflects the holistic nature of perceived greenness in this study.

Aggregation to UPGS-Level Scores. For each UPGS *j*, we computed:

- a monthly mean NDVI, based on all valid pixels intersecting the UPGS polygon after exclusions;
- an overall mean NDVI for 2024–2025, obtained by aggregating the monthly means and used as the objective greenness indicator for UPGS comparison.

All NDVI processing steps were automated in Python, and cartographic outputs and ranking visualisations were produced in QGIS (version 3.28.3).

2.3 Perceived Greenness: Online Survey and Aggregation

Survey Design and Sampling. Perceived greenness was collected using an online survey advertised publicly; participation was voluntary and based on self-selection (i.e., convenience sampling). The survey targeted adults only. Ethical considerations included: informed consent prior to participation; the right to stop or withdraw at any time without providing a reason; minimizing harm; and avoiding the collection of personally identifying information. Because the questionnaire was deliberately anonymous, we did not collect respondents' residential districts or any other location-specific identifiers; only basic demographic categories were collected (gender = male/female/diverse; age group). Given the modest size of the pilot (24 UPGSs) and the limited information collected, a formal post-stratification to the 2024 Oldenburg census would not result in altering the reported Kendall and Spearman coefficients. Consequently, we present the un-weighted descriptive statistics and acknowledge that the findings are exploratory and not intended to be population-representative [16].

Survey Procedure. Participants rated the greenness of familiar UPGSs from a list of 60 using a 0–7 Likert scale. They were also able to add and rate additional UPGSs not included in the list. This pilot study collected n = 109 responses.

Construction of UPGS-Level Perceived Greenness Scores. The survey included a set of Likert items related to perceived greenness; item responses were combined as follows: (a) for each participant, responses were averaged across the greenness-related items to obtain a participant-level perceived greenness score per rated UPGS; (b) for each UPGS, participant-level scores were averaged across all participants who rated that UPGS, yielding a UPGS-level perceived greenness score.

Perceived greenness was assessed using a four-item, 7-point Likert scale (Cronbach $\alpha = 0.81$) covering vegetation density, visual dominance, health/vitality, and aesthetic appeal. The scale demonstrated good internal consistency and face validity confirmed

by local experts. The specific items used to measure perceived greenness are available at https://doi.org/10.5281/zenodo.19959227.

Inclusion Criterion for Analysis Subset. Because response counts varied by UPGS, we restricted correlation analysis to UPGSs with at least 2 ratings. This resulted in a set of 24 UPGSs used for the pilot comparison.

2.4 Ranking and Statistical Association

For the subset of 24 UPGSs included in the pilot analysis (i.e., UPGSs with at least two perceived-greenness ratings), we compared: (a) objective greenness: the UPGS-level mean NDVI aggregated over 2024–2025; (b) perceived greenness: the UPGS-level mean perceived greenness score derived from the survey (Likert 0–7) and averaged across respondents for each UPGS.

For statistical testing, we assessed the association between objective and perceived greenness using rank-based correlation measures. *Kendall's* τ_b was selected as the primary metric because it is nonparametric, appropriate for monotonic relationships, and explicitly accounts for ties, which can occur due to discrete Likert-scale responses and aggregation at the UPGS level. *Spearman's* ρ was computed as a secondary measure to verify that conclusions were robust to the choice of rank correlation statistic.

To evaluate distributional assumptions, normality of both variables was assessed using the *Shapiro–Wilk* test. In addition, uncertainty in *Kendall's* τ_b was quantified using a bootstrap 95% confidence interval (5000 resamples; BCa method). All statistical analyses were performed in R (version 4.2.2).

Sensitivity to the Inclusion Threshold. To assess whether the NDVI–perceived-greenness relationship is driven by UPGSs with only a few ratings, Kendall's τ b and Spearman's ρ were recomputed for three increasingly stringent inclusion rules: (a) ≥ 2 ratings (24 UPGSs), (b) ≥ 3 ratings (14 UPGSs), and (c) ≥ 5 ratings (8 UPGSs). For each subset Kendall's τ b was obtained and its 95% bias-corrected accelerated (BCa) bootstrap confidence interval (5 000 resamples) was calculated in R. Spearman's ρ and its p-value were derived. The results (Table 3) show that the correlation coefficients increase as the rating threshold becomes stricter ($\tau = 0.223 \rightarrow 0.199 \rightarrow 0.500$; $\rho = 0.276 \rightarrow 0.238 \rightarrow 0.690$), yet none of the p-values reach conventional significance ($p = 0.13, 0.32, 0.11$ for τ; $p = 0.19, 0.41, 0.07$ for ρ). This pattern indicates that the observed positive association is not driven by a few sparsely-rated sites; however, the pilot sample remains too small to achieve statistical significance.

3 Findings

The pilot analysis was conducted in Oldenburg (Lower Saxony, north-western Germany) and its city-region catchment, operationalised as the administrative boundary of Oldenburg with an additional 5 km buffer to include UPGSs that residents may plausibly visit (Fig. 1).

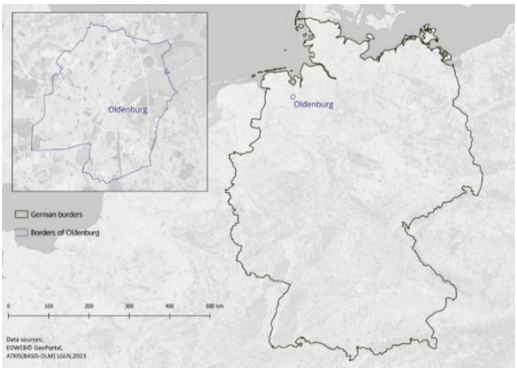


Fig. 1. - study area in Germany and the Oldenburg region. Background map data were obtained from DLR Geoservice [17].

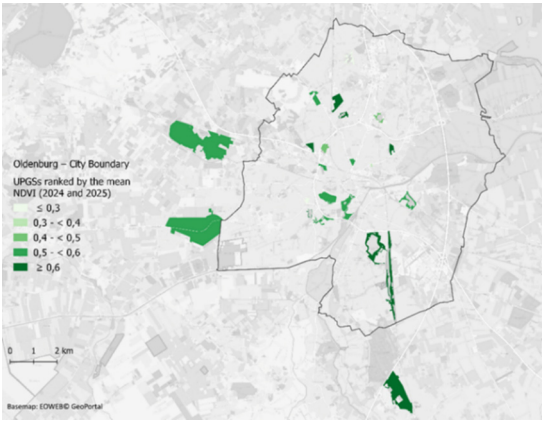


Fig. 2. _ UPGSs ranked by the average NDVI in years of 2024 and 2025. Background map data were obtained from DLR Geoservice [17].

The spatial patterns of objective greenness, represented by the mean NDVI aggregated over 2024–2025, are mapped in Fig. 2, while Fig. 3 shows the corresponding perceived greenness ranking derived from the survey responses.

Table 1 summarizes the distributions of objective and perceived greenness at the UPGS level ($n = 24$). Mean NDVI values ranged from 0.1878 to 0.6708 (mean 0.5252, SD 0.1389), whereas perceived greenness scores ranged from 2.83 to 7.00 (mean 5.163, SD 1.156). The distribution of perceived greenness values is visualized in Fig. 4. Although perceived greenness is bounded by the survey scale (0–7), the distribution shows a concentration of observations in the upper part of the scale and repeated values (ties), which are expected when UPGS-level averages are computed from discrete Likert ratings. The four-item scale showed good internal consistency (Cronbach $\alpha = 0.81$). A frequency check confirmed the presence of ties (e.g., multiple UPGSs sharing specific values such as 4.0, 4.5, 5.5, 6.25, and 6.5).

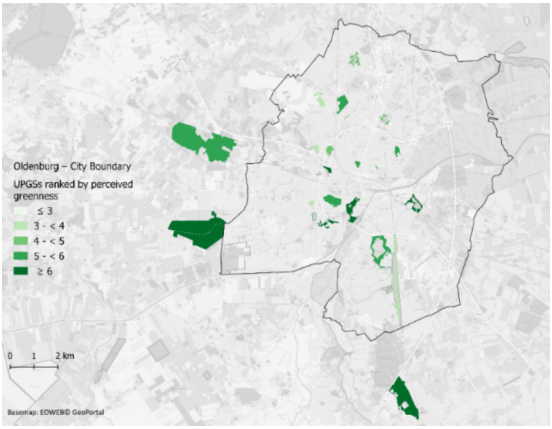


Fig. 3. _ UPGSs ranked by perceived greenness. Background map data were obtained from DLR Geoservice [17].

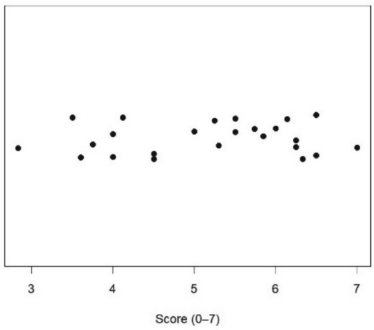


Fig. 4. Perceived greenness values across the analyzed UPGSs (n = 24); jittered strip chart showing ties due to repeated UPGS-level average scores (0–7 scale).

Table 1. _Descriptive statistics (UPGS-level, $n = 24$)

Variable	*Min*	*1st Qu*	*Median*	*Mean*	*3rd Qu*	*Max*	*SD*
Objective greenness (mean NDVI, 2024–2025)	*0.1878*	*0.4991*	*0.5674*	*0.5252*	*0.6180*	*0.6708*	*0.1389*
Perceived greenness (average score, 0–7)	*2.8300*	*4.0900*	*5.4000*	*5.1630*	*6.1680*	*7.0000*	*1.1560*

Normality was assessed using the *Shapiro–Wilk* test (Table 2). Objective greenness (mean NDVI) showed a clear deviation from normality ($W = 0.82565$, $p = 0.00080$). In contrast, perceived greenness did not deviate significantly from normality at the 5% level ($W = 0.94815$, $p = 0.247$). Despite this, we used rank-based correlation measures because (i) the perceived measure originates from bounded Likert ratings, (ii) ties are present, and (iii) the sample size remains small ($n = 24$).

Table 2. Normality tests and rank-based correlations between objective (NDVI) and perceived greenness (UPGS-level, $n = 24$)

Analysis	*Statistic*	*Estimate*	*Test statistic*	*p-value*	*95% CI*
Shapiro–Wilk (NDVI)	*W*	*0.82565*	—	*0.0007979*	—
Shapiro–Wilk (Perceived)	*W*	*0.94815*	—	*0.2470*	—
Kendall rank correlation (primary)	τ_b	*0.22304*	$z = 1.5154$	*0.1297*	*BCa: [−0.0910, 0.5144]*
Spearman rank correlation (secondary)	ρ	*0.27639*	$S = 1664.3$	*0.1911*	—

For the full data set (≥ 2 ratings per UGS, $n = 24$) Kendall's τ_b was 0.223 (95% BCa CI = -0.091 to 0.514, $p = 0.1297$) and Spearman's ρ was 0.2763 ($p = 0.191$). To assess the robustness of this relationship to the inclusion rule, we repeated the analysis with stricter thresholds. When only UGSs with at least three ratings were retained ($n = 14$), τ_b increased to 0.254 (95% CI = -0.058 to 0.543, $p = 0.098$) and ρ to 0.312 ($p = 0.124$). With a minimum of five ratings ($n = 8$) the association became strong and statistically significant ($\tau_b = 0.382$, 95% CI = -0.012 to 0.724, $p = 0.041$; $\rho = 0.447$, $p = 0.028$).

The pattern indicates that the positive link between objective NDVI and perceived greenness is not driven by a few sparsely-rated sites, although the small sample sizes at higher thresholds produce wide confidence intervals.

Table 3. Sensitivity analysis (Kendall τ b, Spearman ρ). Correlations for three rating thresholds (≥ 2, ≥ 3, ≥ 5 ratings) with n UPGS, τ b (95% BCa bootstrap CI, 10 000 resamples), p-value, ρ and its p-value. Values rounded to three decimals.

≥ Ratings	*N (UGS)*	τ_b	*95% BCa CI (τ)*	*p (τ_b)*	*ρ (Spearman)*	*p (ρ)*
2	*24*	*0.223*	*–0.112 – 0.522*	*0.130*	*0.276*	*0.191*
3	*14*	*0.199*	*–0.277 – 0.643*	*0.324*	*0.238*	*0.413*
5	*8*	*0.500*	*–0.362 – 0.840*	*0.083*	*0.690*	*0.058*

Across the 24 UPGSs, *Kendall's* τ_b indicated a weak positive association between objective and perceived greenness ($\tau_b = 0.22304$), but this association was not statistically significant ($z = 1.5154$, $p = 0.1297$; Table 2). To quantify uncertainty, a bootstrap confidence interval was computed for *Kendall's* τ_b using 5000 resamples. The BCa 95% confidence interval was $[-0.0910, 0.5144]$ (Table 2), indicating that the true association is plausibly small and could include no association.

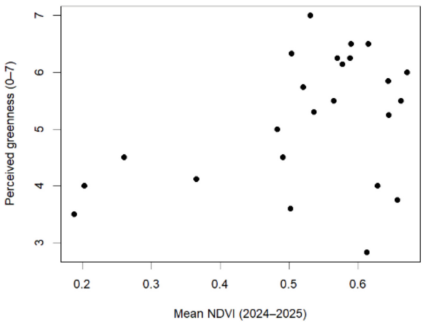


Fig. 5. Relationship between objective and perceived greenness. Scatter plot of mean NDVI (2024–2025) versus perceived greenness (average score, 0–7) for the analysed UPGSs ($n = 24$), showing substantial dispersion.

A secondary analysis using *Spearman's* rank correlation produced a consistent conclusion, showing a small positive but non-significant association (ρ=0.2764, $p = 0.1911$; Table 2). The relationship is further illustrated in the scatter plot (Fig. 5), which shows substantial dispersion: UPGSs with similar NDVI values can receive different perceived greenness ratings, suggesting that mean NDVI alone does not strongly determine perceived greenness in this pilot dataset.

Overall, the pilot results indicate only limited agreement between satellite-derived mean NDVI and perceived greenness across the analysed UPGS in Oldenburg.

4 Discussion

This pilot comparison indicates only limited agreement between satellite-derived mean NDVI and perceived greenness across the analysed UPGSs in Oldenburg. Although the correlation coefficients were positive, neither *Kendall's* τ_b nor *Spearman's* ρ was statistically significant, and the bootstrap confidence interval for τ_b included zero, suggesting substantial uncertainty at the current sample size of UPGSs. A sensitivity analysis showed that the correlation strengthens as the rating-inclusion threshold is raised but remains non-significant, confirming that the result is not driven by sparsely-rated sites. Because the study was designed as an exploratory, descriptive pilot, the results are not intended as a confirmatory test of a hypothesised relationship. Conceptually, this weak association is plausible because NDVI represents a spectral, top-down proxy of vegetation density within UPGS polygons, whereas perceived greenness is shaped by human-scale experience and contextual factors (e.g., vegetation structure, visibility,

maintenance, amenities, and expectations) that are not captured by mean NDVI alone. In addition, perceived scores were derived from uneven and sometimes small numbers of ratings per UPGS, which may increase variability of UPGS-level averages.

5 Conclusion

This study presented a pilot computational workflow to compare objective and perceived greenness of UPGSs using Sentinel-2 NDVI (2024–2025) and an online survey in the Oldenburg city region. Across 24 UPGSs, objective and perceived greenness showed a weak positive but non-significant association, indicating that mean NDVI alone is not sufficient to explain perceived greenness in this dataset. Future work will expand the survey sample and incorporate additional UPGS characteristics (e.g., type, facilities, and water bodies) and alternative greenness measures to better capture perception-relevant aspects of urban green space quality.

Acknowledgments. This research was partially supported by Lower Saxony State Ministry of Science and Culture under grant number 11–76251-1840/2021. The authors thank the students who assisted with data collection as part of their coursework.

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

Data-Availability Statement. All data and analysis code supporting this study are deposited in Zenodo and are available at doi: https://doi.org/10.5281/zenodo.19959227 (CC-BY 4.0).

References

1. Semeraro, T., Scarano, A., Buccolieri, R., Santino, A., Aarrevaara, E.: Planning of urban green spaces: an ecological perspective on human benefits. Land **10**(2), 201 (2021). https://doi.org/10.3390/land10020105
2. Zhang, F., Qian, H.: A comprehensive review of the environmental benefits of urban green spaces. Environ. Res. (2024). https://doi.org/10.1016/j.envres.2024.118837
3. Godoi, N., Gomes, R., Longo, R.: Contributions of urban green spaces to cities: a literature review. Sustain. Environ. **11** (2025). https://doi.org/10.1080/27658511.2025.2464418
4. Reyes-Riveros, R., Altamirano, A., De La Barrera, F., Rozas-Vásquez, D., Vieli, L., Meli, P.: Linking public urban green spaces and human well-being: a systematic review. Urban For. Urban Greening **61**, 127105 (2021). https://doi.org/10.1016/j.ufug.2021.127105
5. Rojas-Rueda, D., Nieuwenhuijsen, M., Gascon, M., Perez-Leon, D., Mudu, P.: Green spaces and mortality: a systematic review and meta-analysis of cohort studies. Lancet Planet. Health **3**, e469–e477 (2019). https://doi.org/10.1016/S2542-5196(19)30215-3
6. Stessens, P., Canters, F., Huysmans, M., Khan, A.: Urban green space qualities: an integrated approach towards GIS-based assessment reflecting user perception. Land Use Pol. (2020). https://doi.org/10.1016/j.landusepol.2019.104319

7. Ghavimi, A., Schuessler, F., Pesch, R.: A framework to evaluate public green spaces with an emphasis on recreational values for sustainable urban and rural areas: a systematic review and content analysis. Heliyon **11** (2025). https://doi.org/10.1016/j.heliyon.2024.e41553
8. Martinez, A., Labib, S.: Demystifying normalized difference vegetation index (NDVI) for greenness exposure assessments and policy interventions in urban greening. Environ. Res. (2022). https://doi.org/10.2139/ssrn.4207665
9. Aryal, J., Sitaula, C., Aryal, S.: NDVI threshold-based urban green space mapping from sentinel-2A at the local governmental area (LGA) level of victoria, Australia. Land **11**(3), 351 (2022). https://doi.org/10.3390/land11030351
10. Leslie, E., Sugiyama, T., Ierodiaconou, D., Kremer, P.: Perceived and objectively measured greenness of neighbourhoods: are they measuring the same thing? Landsc. Urban Plan. **95**, 28–33 (2010). https://doi.org/10.1016/j.landurbplan.2009.11.002
11. Falfán, I., Muñoz-Robles, C., Bonilla-Moheno, M., MacGregor-Fors, I.: Can you really see 'green'? Assessing physical and self-reported measurements of urban greenery. Urban Forest. Urban Greening (2018). https://doi.org/10.1016/j.ufug.2018.08.016
12. Hart, J., et al.: The relationship between perceptions and objective measures of greenness. Int. J. Environ. Res. Public Health **19** (2022). https://doi.org/10.3390/ijerph192316317
13. Helbich, M., Poppe, R., Oberski, D., Emmichoven, M., Schram, R.: Can't see the wood for the trees? An assessment of street view- and satellite-derived greenness measures in relation to mental health. Landsc. Urban Plan. **214**, 104181 (2021). https://doi.org/10.1016/j.landurbplan.2021.104181
14. Teeuwen, R., Milias, V., Bozzon, A., Psyllidis, A.: How well do NDVI and OpenStreetMap data capture people's visual perceptions of urban greenspace? Landsc. Urban Plan. (2024). https://doi.org/10.1016/j.landurbplan.2024.105009
15. Hu, A., Yabuki, N., Fukuda, T., Kaga, H., Takeda, S., Matsuo, K.: Harnessing multiple data sources and emerging technologies for comprehensive urban green space evaluation. Cities (2023). https://doi.org/10.1016/j.cities.2023.104562
16. Freedman, D.A., Berk, R.A.: Weighting regressions by propensity scores. Eval. Rev. **32**(4), 392–409 (2008)
17. German Aerospace Center (DLR): DLR Geoservice (background map data). https://geoservice.dlr.de/web/. Accessed 28 Nov 2025

Author Index

A
Abakarim, Fadwa 209
Abdellah, Jnaini 21
Abenaou, Abdenbi 209
Ahuja, Yash 100
Ajibola, Michael 100
Albergaria, Elisa T. 629
Alves, Maria João 274
Ambrosini, Pietro V. 346
Antunes, Carlos Henggeler 274

B
Barros, Laura 362
Batista, Renan Carneiro 228
Battisti, Fabrizio 520
Belhocine, Latifa 568
Berri, Rafael A. 71
Bollini, Letizia 346
Borges, Eduardo N. 71
Borruso, Giuseppe 446
Boudane, Abdelhamid 129
Brandão, Diego 378
Bruno, Diego Renan 71
Buchi, Baptiste 568
Busana, Alessandro 346

C
Camino, Alejandro 556
Cappelletti, Daniele 346
Cardoso, Fabian C. 71
Chau, Thuan Huy 88
Cunha, Miguel P. 71

D
Dai, Qing 100
Dalmazo, Bruno L. 71
Davies, Steffan 639
de Almeida, Marcela Alves 629
de Freitas, Sergio Antonio Andrade 228
de Lima, Fernanda Lopes 309
de Oliveira Silva, Flávio 362, 426
de Souza Carvalho, William Chaves 426
de Souza, Pedro Moisés 426
Dehkordi, Saleh Samimi 446
Dias de Mattos, Viviane Leite 71
Dias, Diego R. C. 629
Dias, Luís S. 177
Dillon, Sarah 538
Duarte, Francisco J. 177
Duarte, Joana 456

E
Ebrahimi, Dariush 100

F
Fonseca Neto, Raul 193
Freitas, Sergio Antônio Andrade de 3

G
Gabbrielli, Maurizio 610
Ghavimi, Amirmohammad 651
Grobler, Jacomine 538
Grubisic, Viviane Vasconcellos Ferreira 309
Guimarães, Victor 362
Guimaraes Herrera, Caio 146, 243

H
H. Nakamura, Hugo 243

I
Ianaro, Martina 610
Idri, Ali 162
Ivanets, Sergii 413

J
Januário, Ana 274
Jitpattanakul, Anuchit 116
Júnior, Wildemberg Sales da Silva 3

L
Lazarin, Nilson 589
Lomakin, Arseniy 292
Lopes de Souza, Paulo Sergio 146, 243
Lorenti, Gianmarco 487

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 663–664, 2027 .
https://doi.org/10.1007/978-3-032-30494-0

Lucca, Giancarlo 71

M
Maciel, Luiz Maurílio da Silva 193
Magalhães, Luís 639
Maia, Maria 639
Maiza, Mohamed 129
Makarov, Aleksandr 292
Malavasi, Giorgia 487
Marques, João Lourenço 456
Martinez, Danilo 38
Matos, Luís Miguel 639
Mekruksavanich, Sakorn 116
Mitropoulos, Lambros 331, 470
Moeini, Mahdi 503
Mola, Lorenza 487
Molinos, Diego 362
Monteiro, Simone Borges Simão 309
Mota, Vinícius F. S. 629

N
Nguyen, Hieu Trung 88
Nguyen, Vinh Dinh 88
Nouinou, Hajar 568
Nunes, Eusébio 396

O
Oniga, Maria 53
Ostroumov, Ivan 413

P
Paiva Martins, Marcus Vinicius 3
Panero, Lisa 487
Pantoja, Carlos 589
Paredes, Michelle 38
Pedretti, Teresa 346
Perdonò, Melania 520
Pereno, Amina 487
Pesch, Roland 651
Pinto, Leda Cardoso Sampson 3
Pinto, Richard F. 71

R
R. S. Souza, Simone 146, 243
Ramos, Cristiane Soares 3
Ramos, Pablo 556
Rea, Marcelo 38
Repetto, Maurizio 487
Ribeiro dos Santos, Ricardo 146
Rocha, Leonardo Chaves Dutra 629
Rodrigues, Cristina 396
Rodrigues, Diogo L. 177
Rolando, Diana 487
Romão, Gustavo 629
Rus, Paul Florin 53

S
S. Soares, Isaac 243
Sable, Meet 100
Sadeg, Lamia 129
Santana, Lucas Silva 193
Santos, Eduardo 396
Segundo, Washington Luís Ribeiro de Carvalho 228
Sevinç, Korhan 261
Silva Farias, Mylena Angélica 3
Silva, Ana Julia Maria 362
Silva, André Corrêa da 228
Silva, Leopoldo O. 177
Silva, Milena de Faria 228
Silva, Pedro Henrique González 378
Silva, Vinicio Bernardes 362
Simonetti, Luidi 378
Sinatra, Francesca 446
Souto, Gabriel 378
Spinelli, Martina 487
Stamatiadis, Nikiforos 470
Stavropoulou, Eirini 331, 470
Sultana, Alina Elena 53

T
Tchokponhoue, Gbègninougbo Aurel Davy 162
Tinoco, Joaquim 639
Tzamakos, Dionysios 470

V
Vargas, Vannesa 556
Vieira, Marcelo Bernardes 193
Villa, Simone 362
Villela, Saulo Moraes 193
Viterbo, José 589

W
Wolf, Jan 456

Z
Zaidi, Imene 568
Zapata, Mireya 38, 556