

Frontiers
in
Artificial
Intelligence
and
Applications

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

**Proceedings of the 24th International
Conference of the Catalan Association
for Artificial Intelligence**

Edited by
Atia Cortés
Francisco Grimaldo
Tommaso Flaminio



IOS Press

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Artificial intelligence has become an integral part of all our lives. Development is rapid in this exciting and far-reaching field, and keeping up to date with the latest research and innovation is crucial to all those working with the technology.

This book presents the proceedings of the 24th edition of CCIA, the International Conference of the Catalan Association for Artificial Intelligence, held in Sitges, Spain, from 19 – 21 October 2022. This annual event serves as a meeting point not only for researchers in AI from the Catalan speaking territories (southern France, Catalonia, Valencia, the Balearic Islands and Alghero in Italy) but for researchers from around the world. The programme committee received 59 submissions, from which the 26 long papers and 23 short papers selected for presentation at the conference by the 62 experts who make up the committee are included here. The book is divided into the following sections: combinatorial problem solving and logics for artificial intelligence; sentiment analysis and text analysis; data science, recommender systems and decision support systems; machine learning; computer vision; and explainability and argumentation. This book also includes an abstract of the invited talk given by Prof. Fosca Giannotti.

Providing a comprehensive overview of research and development, this book will be of interest to all those working in the field of Artificial Intelligence.



ISBN 978-1-64368-326-3 (print)

ISBN 978-1-64368-327-0 (online)

ISSN 0922-6389 (print)

ISSN 1879-8314 (online)

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including ‘Information Modelling and Knowledge Bases’ and ‘Knowledge-Based Intelligent Engineering Systems’. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.

The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Series Editors:

Nicola Guarino, Pascal Hitzler, Joost N. Kok, Jiming Liu, Ramon López de Mántaras,
Riichiro Mizoguchi, Mark Musen, Sankar K. Pal, Ning Zhong

Volume 356

Recently published in this series

- Vol. 355. H. Fujita, Y. Watanobe and T. Azumi (Eds.), New Trends in Intelligent Software Methodologies, Tools and Techniques – Proceedings of the 21st International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_22)
- Vol. 354. S. Schlobach, M. Pérez-Ortiz and M. Tielman (Eds.), HHAI2022: Augmenting Human Intellect – Proceedings of the First International Conference on Hybrid Human-Artificial Intelligence
- Vol. 353. F. Toni, S. Polberg, R. Booth, M. Caminada and H. Kido (Eds.), Computational Models of Argument – Proceedings of COMMA 2022
- Vol. 352. A.J. Tallón-Ballesteros (Ed.), Modern Management based on Big Data III – Proceedings of MMBD 2022
- Vol. 351. A. Passerini and T. Schiex (Eds.), PAIS 2022 – 11th Conference on Prestigious Applications of Artificial Intelligence, 25 July 2022, Vienna, Austria (co-located with IJCAI-ECAI 2022)
- Vol. 350. P. Morettin, Learning and Reasoning in Hybrid Structured Spaces
- Vol. 349. A. De Filippo, Hybrid Offline/Online Methods for Optimization Under Uncertainty
- Vol. 348. G. Cima, Abstraction in Ontology-based Data Management

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Artificial Intelligence Research and Development

Proceedings of the 24th International Conference of the Catalan Association for Artificial Intelligence

Edited by

Atia Cortés

Barcelona Supercomputing Center, Spain

Francisco Grimaldo

Universitat de València, Spain

and

Tommaso Flaminio

Institut d'Investigació en Intel·ligència Artificial, Spain



IOS Press

Amsterdam • Berlin • Washington, DC

© 2022 The authors and IOS Press.

This book is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

ISBN 978-1-64368-326-3 (print)

ISBN 978-1-64368-327-0 (online)

Library of Congress Control Number: 2022946684

doi: 10.3233/FAIA356

Publisher

IOS Press BV

Nieuwe Hemweg 6B

1013 BG Amsterdam

Netherlands

fax: +31 20 687 0019

e-mail: order@iospress.nl

For book sales in the USA and Canada:

IOS Press, Inc.

6751 Tepper Drive

Clifton, VA 20124

USA

Tel.: +1 703 830 6300

Fax: +1 703 830 2300

sales@iospress.com

LEGAL NOTICE

The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Preface

The International Conference of the Catalan Association for Artificial Intelligence (CCIA) is an event which serves as a meeting point, not only for researchers in Artificial Intelligence based in the area of the Catalan speaking territories (southern France, Catalonia, Valencia, the Balearic Islands and Alghero in Italy), but also for researchers around the world.

This book constitutes the proceedings of the 24th edition of the CCIA, held in Sitges, in October 2022. Previous editions of the CCIA were held in Tarragona (1998), Girona (1999), Vilanova i la Geltrú (2000), Barcelona (2001, 2004, 2014, 2016), Castelló de la Plana (2002), Mallorca (2003), Alghero (Sardinia) (2005), Perpignan' (France) (2006), Andorra (2007), Sant Martí d'Empúries (2008), Cardona (2009), L'Espluga de Francolí (2010), Lleida (2011), Alacant (2012), Vic (2013), València (2015), Deltebre (2017), Roses (2018), Colònia de Sant Jordi (2019) and Lleida (2021). CCIA was cancelled in 2020 due to the restrictions caused by the COVID-19 outbreak.

The 26 long papers and the 23 short papers presented in this volume were carefully reviewed and selected from 59 submissions. This reviewing process was made possible thanks to the 62 artificial intelligence experts who make up the programme committee. We especially thank them for their efforts in this task, and would also like to express our appreciation for the work of the authors of the 59 submissions.

The accepted papers deal with all aspects of artificial intelligence, including combinatorial problem solving and logics for artificial intelligence, sentiment analysis and text analysis, data science, recommender systems and decision support systems, machine learning, computer vision, and explainability and argumentation. This book of proceedings also includes the abstract of the invited talk, given by Prof. Fosca Giannotti.

We would like to express our sincere gratitude to the Catalan Association for Artificial Intelligence (ACIA), the Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), the Universitat de València (UV) and the Barcelona Supercomputing Center (BSC) for their support.

Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), October 2022

Tommaso Flaminio, Institut d'Investigació en Intel·ligència Artificial
Francisco Grimaldo, Universitat de València
Atia Cortés, Barcelona Supercomputing Center

This page intentionally left blank

About the Conference

The CCIA 2022 conference was organized by the Associació Catalana d'Intel·ligència Artificial (ACIA), the Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), the Barcelona Supercomputing Center (BSC) and the Universitat de València (UV)

General Chair

Francisco Grimaldo, Universitat de València

Scientific Chair

Atia Cortés, Barcelona Supercomputing Center

Local Organizing Chairs

Tommaso Flaminio, Institut d'Investigació en Intel·ligència Artificial

Local Organizing Committee

Eva Armengol, Institut d'Investigació en Intel·ligència Artificial

Lluís Godó, Institut d'Investigació en Intel·ligència Artificial

Josep Puyol, Institut d'Investigació en Intel·ligència Artificial

Marco Schorlemmer, Institut d'Investigació en Intel·ligència Artificial

Sara Ugolini, Institut d'Investigació en Intel·ligència Artificial

Amanda Vidal, Institut d'Investigació en Intel·ligència Artificial

Scientific Committee

Isabel Aguiló (UIB)

René Alquézar (UPC)

Sergio Alvarez-Napagao (BSC)

Cecilio Angulo (UPC)

Javier Antich (UIB)

Josep Lluís Arcos (IIIA-CSIC)

Eva Armengol (IIIA-CSIC)

Javier Béjar (UPC)

Ester Bernadó (Tecnocampus)

Francisco Bonnín (SRV-UIB)

Josep Casas-Roma (UOC)

Gustavo Casañ (RobIn Lab, UJI)

Jesús Cerquides (IIIA-CSIC)

Hubie Chen (University of London)

Jordi Coll (Aix-Marseille Université)

Ulises Cortés (UPC)

Joan Espasa (University of St. Andrews)

Zoe Falomir (UJI)

Francesc J. Ferri (UV)

Emilio García (UIB)

Dario Garcia-Gasulla (BSC)

Ricard Gavaldà (UPC)

Héctor Geffner (ICREA-UPF)

Lluís Godó (IIIA-CSIC)

Elisabet Golobardes (URL)

Manuel González (UIB)

José M. Iñesta (UA)

Anders Jonsson (UPF)

Vicente Julián (UPV)

Jordi Levy (IIIA-CSIC)

Beatriz López (UdG)

Joaquim Meléndez (UdG)

Pedro Meseguer (IIIA-CSIC)

Antonio Moreno (URV)

Lledó Museros (UJI)	Maria Salamó (UB)
Ángela Nebot (UPC)	Miquel Sàncchez (UPC)
Carles Noguera (UTIA-CAS)	Marco Schorlemmer (IIIA-CSIC)
Enric Plaza (IIIA-CSIC)	Vicenç Torra
Ferran Padrós (UOC)	Carme Torras (IRI-CSIC-UPC)
Josep Puyol (IIIA-CSIC)	Joaquín Torres (UJI)
David Riaño (URV)	Aïda Valls (URV)
Juan Vicente Riera (UIB)	Xavier Varona (UIB)
Andrea Rizzoli (IDSIA, SUPSI)	Javier Vázquez-Salceda (UPC)
Juan Antonio Rodríguez (IIIA-CSIC)	Alfredo Vellido (UPC)
Jordi Sabater (IIIA-CSIC)	Franz Wotawa (TU Graz)

Organizing Institutions



Contents

Preface <i>Tommaso Flaminio, Francisco Grimaldo and Atia Cortés</i>	v
About the Conference	vii
Invited Talk	
Explainable Machine Learning for Trustworthy AI <i>Fosca Giannotti</i>	3
Combinatorial Problem Solving and Logics for Artificial Intelligence	
Towards an Implementation of Merging Operators in Many-Valued Logics <i>Vicent Costa and Pilar Dellunde</i>	7
Multi Objective Genetic Algorithm for Optimal Route Selection from a Set of Recommended Touristic Activities <i>Jonathan Ayebakuro Orama, Antonio Moreno and Joan Borras</i>	9
On Conjunctive and Disjunctive Rational Bivariate Aggregation Functions of Degrees (2,1) <i>Isabel Aguiló, Sebastia Massanet and Juan Vicente Riera</i>	13
Approximate and Optimal Solutions for the Bipartite Polarization Problem <i>Teresa Alsinet, Josep Argelich, Ramón Béjar and Santi Martínez</i>	17
Yet Another (Fake) Proof of P=NP <i>Carlos Ansótegui and Jordi Levy</i>	25
A Tableau Calculus for MaxSAT Based on Resolution <i>Shoulin Li, Jordi Coll, Djamal Habet, Chu-Min Li and Felip Manyà</i>	35
Importance-Performance Analysis in Project Portfolio Management Using an IOWA Operator <i>Pietro Fronte, Núria Agell, Marc Torrens and Daniel Brugarolas</i>	45
Sentiment Analysis and Text Analysis	
Streamlining Text Pre-Processing and Metrics Extraction <i>Elena Alvarez-García, Daniel García-Costa and Francisco Grimaldo</i>	55
Drake or Hen? Machine Learning for Gender Identification on Twitter <i>Arnault Gombert and Jesus Cerquides</i>	59

Analysing Food-Porn Images for Users' Engagement in the Food Business <i>V. Casales-Garcia, Z. Falomir, L. Museros, I. Sanz, D.M. Llido and L. Gonzalez-Abril</i>	67
Influence in Social Networks Through Visual Analysis of Image Memes <i>Carles Onielfa, Carles Casacuberta and Sergio Escalera</i>	71
Data Science, Recommender Systems and Decision Support Systems	
Deep Air – A Smart City AI Synthetic Data Digital Twin Solving the Scalability Data Problems <i>Esteve Almirall, Davide Callegaro, Peter Bruins, Mar Santamaría, Pablo Martínez and Ulises Cortés</i>	83
Optimizing Online Time-Series Data Imputation Through Case-Based Reasoning <i>Josep Pascual-Pañach, Miquel Sánchez-Marrè and Miquel Àngel Cugueró-Escofet</i>	87
Synthetic Data for Anonymization in Secure Data Spaces for Federated Learning <i>Cecilio Angulo and Cristóbal Raya</i>	91
Towards Automated Compliance Checking of Building Regulations: smartNorms4BIM <i>Ignacio Huitzil, Marco Schorlemmer, Nardine Osman, Pere Garcia, Josep Coll and Xavier Coll</i>	95
Deep Neural Classification of Darknet Traffic <i>Mahmoud Alimoradi, Mahdieh Zabihimayvan, Arman Daliri, Ryan Sledzik and Reza Sadeghi</i>	105
Enabling Reproducibility in Group Recommender Systems <i>Joaquin Dario Silveira, Maria Salamó and Ludovico Boratto</i>	115
Contextual TV Show Recommendation <i>Paula Gómez Duran and Jordi Vitrà</i>	125
Machine Learning	
Predicting Personalized Quality of Life of an Intellectually Disabled Person Utilizing Machine Learning <i>Gaurav Kumar Yadav, Benigno Moreno Vidales, Sara Dueñas, Mohamed Abdel-Nasser, Hatem A. Rashwan, Domènec Puig and G.C. Nandi</i>	139
The Assessment of Clustering on Weighted Network with R Package <i>clustAnalytics</i> <i>Argimiro Arratia and Martí Renedo-Mirambell</i>	143
Binary Delivery Time Classification and Vehicle's Reallocation Based on Car Variants. SEAT: A Case Study <i>Juan Manuel García Sánchez, Xavier Vilasis Cardona and Alexandre Lerma Martín</i>	147

Feature Engineering and Machine Learning Predictive Quality Models for Friction Stir Welding Defect Prediction in Aerospace Applications <i>Marta Camps, Maddi Etxegarai, Francesc Bonada, William Lacheny, Sylvain Pauleau and Xavier Domingo</i>	151
Efficiency and Reliability Enhancement of High Pressure Die Casting Process Through a Digital Twin <i>Pol Torres, Albert Abio, Raquel Busqué, Albert Brígido, Sylvia Andrea Cruz, Manel Da Silva and Francesc Bonada</i>	155
An Agent-Based Simulation Framework for Firefighters Training <i>Jordi Sabater-Mir, Ignasi Camps-Ortin and Cristian Cozar-Alier</i>	160
Data Driven Predictive Models Based on Artificial Intelligence to Anticipate the Presence of <i>Plasmopara Viticola</i> and <i>Uncinula Necator</i> in Southern European Winegrowing Regions <i>Marta Otero, Luisa Fernanda Velasquez, Boris Basile, Jordi Ricard Onrubia, Alex Josep Pujol and Josep Pijuan</i>	164
Towards and Efficient Algorithm for Computing the Reduced Mutual Information <i>Martí Renedo-Mirambell and Argimiro Arratia</i>	168
Bootstrap-CURE Clustering: An Investigation of Impact of <i>Shrinking</i> on Clustering Performance <i>Ashutosh Karna and Karina Gibert</i>	172
Adapting a Fuzzy Random Forest for Ordinal Multi-Class Classification <i>Jordi Pascual-Fontanilles, Lenka Lhotska, Antonio Moreno and Aida Valls</i>	181
On Flows of Neural Ordinary Differential Equations That Are Solutions of Lotka-Volterra Dynamical Systems <i>Argimiro Arratia, Carlos Ortiz and Marcel Romani</i>	191
Garment Manipulation Dataset for Robot Learning by Demonstration Through a Virtual Reality Framework <i>Arnaud Boix-Granell, Sergi Foix and Carme Torras</i>	199
Long Short-Term Memory to Predict 3D Amino Acids Positions in GPCR Molecular Dynamics <i>Juan Manuel López-Correa, Caroline König and Alfredo Vellido</i>	209
Computer Vision	
Towards Cross-Sites Generalization for Prostate MRI Segmentation to Unseen Data <i>Eddardaa Ben Loussaief and Domenec Puig</i>	221
A Curated Dataset for Crack Image Analysis: Experimental Verification and Future Perspectives <i>Ammar M. Okran, Mohamed Abdel-Nasser, Hatem A. Rashwan and Domenec Puig</i>	225

Reducing the Learning Domain by Using Image Processing to Diagnose COVID-19 from X-Ray Image <i>Maider Abad, Jordi Casas-Roma and Ferran Prados</i>	229
On the Importance of Color Pre-Processing for Object Detection in Submarine Images <i>Jose-Luis Lisani, Ana Belén Petro, Catalina Sbert, Amaya Álvarez-Ellacuría, Ignacio A. Catalán and Miquel Palmer</i>	239
Referenceless Image Quality Assessment Utilizing Deep Transfer-Learned Features <i>Basma Ahmed, Osama A. Omer, Amal Rashed, Domenec Puig and Mohamed Abdel-Nasser</i>	243
Detecting the Area of Bovine Cumulus Oocyte Complexes Using Deep Learning and Semantic Segmentation <i>Georgios Athanasiou, Jesus Cerquides, Annelies Raes, Nima Azari-Dolatabad, Daniel Angel-Velez, Ann Van Soom and Josep-Lluis Arcos</i>	249
Object Segmentation of Cluttered Airborne LiDAR Point Clouds <i>Mariona Carós, Ariadna Just, Santi Seguí and Jordi Vitrià</i>	259
Breast Tumor Classification in Digital Tomosynthesis Based on Deep Learning Radiomics <i>Loay Hassan, Mohamed Abdel-Nasser, Adel Saleh and Domenec Puig</i>	269
Automatic Outdoor Image Geolocation with Focal Modulation Networks <i>Fabio Murgese, Gerard Alcaina, Mehmet Oğuz Müläyim, Jesus Cerquides and Jose Luis Fernandez-Marquez</i>	279
Analyzing the Reliability of Different Machine Radiomics Features Considering Various Segmentation Approaches in Lung Cancer CT Images <i>Maryam Tahmooresi, Mohamed Abdel-Nasser and Domenec Puig</i>	289
Transformer-Based Radiomics for Predicting Breast Tumor Malignancy Score in Ultrasonography <i>Mohamed A. Hassani, Vivek Kumar Singh, Domenec Puig and Mohamed Abdel-Nasser</i>	298
<i>EDBNet: Efficient Dual-Decoder Boosted Network for Eye Retinal Exudates Segmentation</i> <i>Mohammed Yousef Salem Ali, Mohamed Abdel-Nasser, Aida Valls, Marc Baget and Mohammed Jabreel</i>	308
Explainability and Argumentation	
Combining Support and Attack Interactions for Argumentation Based Discussion Analysis <i>Teresa Alsinet, Josep Argelich and Ramón Béjar</i>	321
Focus and Bias: Will It Blend? <i>Anna Arias-Duart, Ferran Parés, Victor Giménez-Ábalos and Dario Garcia-Gasulla</i>	325

An Ethical Conversational Agent to Respectfully Conduct In-Game Surveys <i>Eric Roselló-Marín, Maite Lopez-Sánchez, Inmaculada Rodríguez, Manel Rodríguez-Soto and Juan A. Rodríguez-Aguilar</i>	335
Fuzzy-LORE: A Method for Extracting Local and Counterfactual Explanations Using Fuzzy Decision Trees <i>Najlaa Maaroof, Antonio Moreno, Mohammed Jabreel and Aida Valls</i>	345
Testing Reinforcement Learning Explainability Methods in a Multi-Agent Cooperative Environment <i>Marc Domènec i Vila, Dmitry Gnatyshak, Adrián Tormos and Sergio Alvarez-Napagao</i>	355
Mutual Information Weighing for Probabilistic Movement Primitives <i>Adrià Colomé and Carme Torras</i>	365
Subject Index	369
Author Index	373

This page intentionally left blank

Invited Talk

This page intentionally left blank

Explainable Machine Learning for Trustworthy AI

Fosca GIANNOTTI¹

Scuola Normale Superiore, Pisa (Italy)

Information Science and Technology Institute “A. Faedo” of the National Research Council, Pisa (Italy)

Keywords. Explainable AI, Trustworthy AI, Transparency, Machine Learning, Symbolic AI

Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user’s features into a class or a score without exposing the reasons why. This is problematic not only for the lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. The future of AI lies in enabling people to collaborate with machines to solve complex problems. Like any efficient collaboration, this requires good communication, trust, clarity and understanding.

Explainable AI addresses such challenges and for years different AI communities have studied such topic, leading to different definitions, evaluation protocols, motivations, and results. This lecture provides a reasoned introduction to the work of Explainable AI (XAI) to date, and surveys the literature with a focus on machine learning and symbolic AI related approaches. We motivate the needs of XAI in real-world and large-scale application, while presenting state-of-the-art techniques and best practices, as well as discussing the many open challenges.

¹Corresponding Author: Fosca Giannotti, fosca.giannotti@isti.cnr.it

This page intentionally left blank

Combinatorial Problem Solving and Logics for Artificial Intelligence

This page intentionally left blank

Towards an Implementation of Merging Operators in Many-Valued Logics

Vicent COSTA ^{a,b,1}, Pilar DELLUNDE ^{a,b,c,2}

^aDepartament de Filosofia & Institut d’Història de la Ciència (IHC), Universitat Autònoma de Barcelona,

^bInstitut d’Investigació en Intel·ligència Artificial (IIIA-CSIC)

^cBarcelona Graduate School of Mathematics (BGSMath)

Keywords. Belief merging, Signed logic, Horn fragment, Implementation, Many-valued logics

In this paper, we present work in progress on belief merging operators’ implementation in many-valued logics. Belief merging is an active area of artificial intelligence and cognitive science [1,2]. Commonly, it studies the combination of consistent knowledge bases (possibly mutually inconsistent) to obtain a single consistent knowledge base, expressed in classical propositional logic. Recently, some authors restricted the theoretical study to some fragments of propositional logic with good computational properties such as the Horn fragment, which affords very efficient algorithms [3,4]. Likewise, some implementations have been proposed with practical applications using partial-satisfiability-based merging [5]. For instance, Kareem et al. [6] used belief merging for oral cancer diagnosis, and Pozos-Parra et al. [7] introduced Merginator, a belief merging tool for consensus decision making.

Traditionally, the logical approach has proposed a set of postulates that a merging operator has to meet to be considered rational. In previous works [8] we studied the logical postulates and theoretical properties of merging operators in the Horn fragment of many-valued logics. In particular, the formalism we used is signed logic. Signed logic was introduced as a generic treatment of many-valued logics [9], and its fundamental underlying idea is attaching a sign or label to an atomic formula to generalize the classical notion of a literal.

A *knowledge base* can be expressed by a signed Horn formula, defining a *profile* as a nonempty finite multiset of consistent but not necessarily mutually consistent knowledge bases. A profile can be viewed as a multiset of agents, represented by their sets of beliefs. Then, merging is finding a ranking that approximates the individual rankings as best as possible. These are the main results we will present:

1. A characterization of the class of models of a Horn signed formula in terms of the closure of this class under certain operations.
2. A sufficient condition for a signed Horn merging operator to satisfy the logical IC-postulates ([10,11]), showing that Horn merging can be seen as an aggregation problem on rankings of signed interpretations.

¹E-mail: vicent.costa@protonmail.com.

²E-mail: pilar.dellunde@uab.cat.

3. An example of a signed Horn merging operator, defined by giving a notion of distance between signed interpretations (to induce preorders for each knowledge base) and using an aggregation function (to combine the individual rankings into a final preorder for the profile).
4. An example of an implementation of the belief merging process in the Horn fragment of signed logic. Two parts complete this task: the first one consists in implementing the transformation of signed regular Horn formulas into classical Horn formulas; the other one is defining a Horn merging operator meeting the rationality postulates of belief merging [10,11]. This operator is defined by giving a notion of distance between signed interpretations (to induce preorders for each knowledge base) and using an aggregation function (to combine the individual rankings into a final preorder for the profile). We consider a running example of a simple review conference process written in Python.

Acknowledgments

Vicent Costa is a *Margarita Salas Researcher* (European Union - NextGenerationEU). This work is also partially funded by the H2020-MSCA-RISE-2020 project MOSAIC (101007627) and the project ISINC (PID2019-111544GB-C21).

References

- [1] C. Baral, S. Kraus, J. Minker, Combining multiple knowledge bases, *IEEE Trans. Knowl. Data Eng.* 3 (2) (1991) 208–220.
- [2] P. Z. Revesz, On the semantics of theory change: Arbitration between old and new information, in: C. Beeri (Ed.), *Proceedings of the Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, ACM Press, 1993, pp. 71–82.
- [3] N. Creignou, O. Papini, S. Rümmele, S. Woltran, Belief merging within fragments of propositional logic, *ACM Trans. Comput. Log.* 17 (3) (2016) 20:1–20:28.
- [4] A. Haret, S. Rümmele, S. Woltran, Merging in the Horn fragment, *ACM Trans. Comput. Log.* 18 (1) (2017) 6:1–6:32.
- [5] P. P. Parra, V. B. Macías, Partial satisfiability-based merging, in: A. F. Gelbukh, A. F. K. Morales (Eds.), *MICAI 2007: Advances in Artificial Intelligence*, 6th Mexican International Conference on Artificial Intelligence, Vol. 4827 of Lecture Notes in Computer Science, Springer, 2007, pp. 225–235.
- [6] S. A. Kareem, P. P. Parra, N. Wilson, An application of belief merging for the diagnosis of oral cancer, *Appl. Soft Comput.* 61 (2017) 1105–1112.
- [7] M. del Pilar Pozos Parra, O. Chávez-Bosquez, K. McAreavey, Merginator: A belief merging tool for consensus support, *J. Intell. Fuzzy Syst.* 34 (5) (2018) 3199–3210.
- [8] P. Dellunde, A characterization of belief merging operators in the regular Horn fragment of signed logic, in: V. Torra, Y. Narukawa, J. Nin, N. Agell (Eds.), *Modeling Decisions for Artificial Intelligence - 17th International Conference, MDAI 2020*, Vol. 12256 of Lecture Notes in Computer Science, Springer, 2020, pp. 3–15.
- [9] R. Hähnle, *Automated deduction in multiple-valued logics*, Vol. 10 of International series of monographs on computer science, Oxford University Press, 1994.
- [10] S. Konieczny, R. P. Pérez, Merging information under constraints: A logical framework, *J. Log. Comput.* 12 (5) (2002) 773–808.
- [11] S. Konieczny, R. P. Pérez, Logic based merging, *J. Philos. Log.* 40 (2) (2011) 239–270.

Multi Objective Genetic Algorithm for Optimal Route Selection from a Set of Recommended Touristic Activities

Jonathan Ayebakuro ORAMA^{a,1}, Antonio MORENO^b and Joan BORRAS^a

^a*Eurecat, Centre Tecnològic de Catalunya, Vila-Seca, Spain*

^b*ITAKA Research Group, Universitat Rovira i Virgili, Tarragona, Spain*

ORCID ID: Jonathan Ayebakuro Orama <https://orcid.org/0000-0002-2622-3224>,

Antonio Moreno <https://orcid.org/0000-0003-3945-2314>, Joan Borrás

<https://orcid.org/0000-0002-3894-4010>

Abstract. It is a known fact that the order in which touristic activities are experienced plays a role in how enjoyable they are. This is the reason why tourists prefer to book carefully prepared day tours on arrival to a new destination, as they allow them to see the essence of the destination while traversing scenic routes. Tours are great, but they are expensive, do not allow room for personal exploration, and are built as a one-size-fits-all which does not consider the individual preferences of the tourist. In contrast, it is possible to make an optimal selection and ordering of touristic activities from a larger set of possibilities that match a tourist's personal preferences, balancing important aspects like diversity, spatial proximity, or degree of interest on popular places. We propose a multi-objective genetic algorithm that uses a weighted averaging operator to balance four diverse objective functions crafted to maintain diversity, proximity, interest on popularity, and cultural preference. The system has been evaluated against four baseline algorithms and found to perform significantly better for the specified purpose.

Keywords. Multi-Objective Genetic Algorithm, Travel Route Optimization, Weighted Average Objective Balancing

1. Introduction

Tourists are aptly named for their interest in touring new destinations to experience new cultures, cuisines, etc. Effectively touring a new destination requires prior information on attractions to visit, routes to take, or local cuisines to sample, which can only be gotten from destination marketers, who design purchasable guided tours which allow tourists to experience the destination without further planning. This is beneficial to both the destination marketer and the tourist, as the latter can relax and enjoy a leisure trip while the former can use the opportunity to increase the visibility of certain attractions while profiting from sales. However, these tours lack personalisation and don't allow for much exploration.

¹Corresponding Author: Jonathan Ayebakuro Orama, Eurecat, Centre Tecnològic de Catalunya, C/ Joanot Martorell, 15, 43480 Vila-Seca, Spain; E-mail: jonathan.orama@eurecat.org

Quite a bit of research has gone into developing algorithms for building and recommending personalised tours [1]. One of them is the *Genetic Algorithm* (GA), which is a heuristic search algorithm that is relatively fast at finding a good solution for a combinatorial optimization problem. A variant called *Multi-Objective Genetic Algorithm* (MOGA) is perfect for building personalised tours because solutions are optimized to meet multiple objectives. Notable works that use MOGA for tour recommendation start by defining constraints relevant to tourists (e.g. time, distance, budget), and then they evaluate possible solutions from a list of Points of Interest (POIs) based on these constraints while ensuring solutions match tourist preferences [2,3,4]. Our proposed method extends our previous work [5], which recommends a set of POIs that match the user's preference considering their relatedness. It addresses the problem of selecting and ordering a subset of the recommendable POIs to minimize the distance between them and maximize their diversity while maintaining the ratio of popular POIs and types of POIs preferred by the user. These four criteria are formulated into objective functions which are balanced using a weighting vector.

2. Methodology

Our previous recommender system utilises trace data provided on Twitter to build user profiles that capture the interests and travel habits of tourists for recommending a set of related POIs considering their popularity and category. A full description of data collection, preprocessing and the recommendation process can be found at [5]. The output POIs from the recommender system are used in the MOGA to find the solution that fits better the requirements.

The problem is formulated as a combinatorial optimization with the objective of finding an optimal combination of objects from a finite set of objects. In this case, we are looking for a combination of five POIs from a set of ten recommended POIs. Let $R = \{r_1, \dots, r_n\}$ be the set of recommended POIs, and $S = \{s_1, \dots, s_k\} \equiv \{(r_{n1}, \dots, r_{n5}), \dots, (r_{m1}, \dots, r_{m5})\}$ be the search space of possible solutions. The goal of the MOGA is to pick a solution s_i from S that best balances the objectives.

Our proposed algorithm comprises four objective functions that are combined using weighted averaging and a predefined weighting vector to form the fitness function. The objectives are defined as follows:

- **Proximity:** This objective ensures that the mean distance between adjacent POIs is minimized.
- **Diversity:** This objective ensures that the solution contains a diverse set of POIs. POIs are categorised in [5] using a hierarchical structure called an *Activity Tree*. As such, POIs are diverse when they share less categories and subcategories in their path.
- **Popularity:** This objective ensures that the ratio of popular POIs in the solution matches the tourist's degree of interest in popular POIs (computed in [5] as the percentage of tweets sent from popular spots).
- **Preference:** This objective ensures that the POIs in the solution are interesting to the tourist. Relevant POIs are categorised under types of activities frequently visited by the tourist (obtained with an analysis of the places from which the user has tweeted, [5]).

The main steps of the MOGA are the following:

- Step 1.** Generate randomly an initial population, which is a subset of S of size 150.
- Step 2.** Calculate the objective functions and aggregate them using weighted averaging to get the fitness values of all members of population.
- Step 3.** Select two solutions with a good fitness value (i.e. better objective function score) from the population as *parents*, then *crossover* them to form two *children* according to a crossover rate (60%), and then *mutate* children by swapping the positions of POIs according to a mutation rate (1%). If children are fitter than their parents they are added to the new population, otherwise parents are added to the new population (*weak parent replacement*).
- Step 4.** Repeat step 3 until the new population is up to size 150, signaling the completion of generation 1. Cache the best solution in the new population.
- Step 5.** Repeat steps 2 to 4, to move through generations replacing the cached best solution with any better solution. If there is no better solution for 20 generations or 100 generations are completed, stop and return the best solution.

3. Experiments and Results

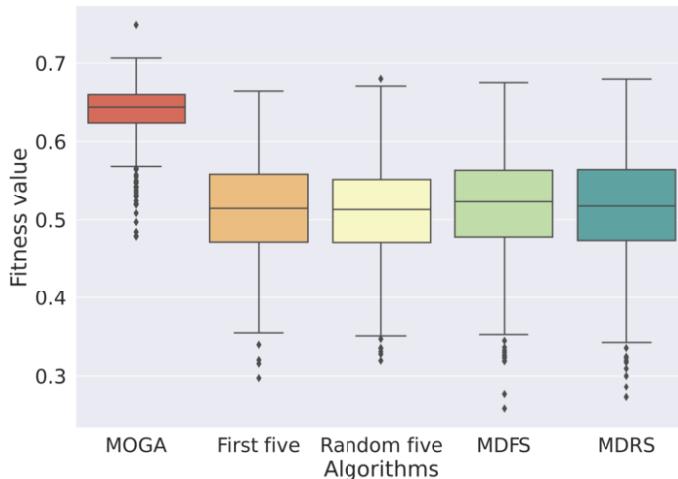
For the experiments, the dataset consisted of 1140 Twitter users with 10 recommended POIs each suggested by our recommender system, which extracts the user's interest in activities, popular or unpopular attractions, and their touring preferences from their tweets in order to make recommendations [5]. The weighting vector [proximity:0.4, diversity:0.3, popularity:0.1, preference:0.2] was used in the fitness function. Each user's POIs were passed through the MOGA and the fitness of the results was compared against the following four baseline algorithms:

1. **First five:** Select the first 5 recommended POIs as the solution (i.e. the ones considered better by the recommender).
2. **Random five:** Select randomly 5 of the 10 recommended POIs as solution.
3. **Minimize distance first start (MDFS):** Select 5 of the 10 recommended POIs that minimize the distance between adjacent POIs using a greedy algorithm with the following steps:
 - (a) Add the *first* POI from the 10 recommended POIs to the initial solution.
 - (b) Add a POI in front of the last POI in solution or behind the first POI of the solution with the least distance to travel.
 - (c) Repeat step b until the solution contains 5 POIs.
4. **Minimize distance random start (MDRS):** This algorithm is the same as MDFS but the first POI added to the initial solution is picked randomly.

Table 1 showcases the fitness value of our MOGA against all baseline algorithms, highlighting the minimum, maximum and mean values across all 1140 users. MOGA performs significantly better than all baseline algorithms. This result is further enforced in Figure 1, which shows a box plot of MOGA and all baseline algorithms. The MOGA results outperform the baseline algorithms, as the lower quartile of MOGA sits above the upper quartiles of the rest. A sequential check of all possible solutions is feasible in this case due to a small search space of 30,240 options, but a test showed the MOGA to be 27% faster with a runtime of 7 minutes per case compared to 11 when searched sequentially, while obtaining the same optimal result.

Table 1. Minimum, maximum and mean fitness values for all algorithms.

Fitness value	MOGA	First five	Random five	MDFS	MDRS
Minimum	0.479	0.297	0.319	0.258	0.273
Maximum	0.748	0.665	0.680	0.676	0.680
Mean	0.639	0.512	0.510	0.519	0.516

**Figure 1.** Box plots of fitness values for all algorithms.

4. Conclusion

In this paper we have proposed a multi-objective genetic algorithm for selecting and ordering a fixed set of POIs from a larger set of recommended POIs. It works by balancing four objective functions, *proximity*, *diversity*, *popularity*, and *preference* using weighted averaging. Experiments show our MOGA to outperform certain baseline algorithms.

References

- [1] Lim KH, Chan J, Karunasekera S, Leckie C. Tour recommendation and trip planning using location-based social media: a survey. *Knowl Inf Syst*. 2019 Dec; 60:1247–1275, doi:10.1007/s10115-018-1297-4.
- [2] Yochum P, Chang L, Gu T, Zhu M. An Adaptive Genetic Algorithm for Personalized Itinerary Planning. *IEEE Access*. vol. 2020 Apr; 8:88147-88157, doi:10.1109/ACCESS.2020.2990916.
- [3] Zheng X, Luo Y, Sun L, Yu Q, Zhang J, Chen S. A Novel Multi-Objective and Multi-Constraint Route Recommendation Method Based on Crowd Sensing. *Appl Sci*. 2021 Nov; 11(21):10497, doi:10.3390/app112110497.
- [4] Yuan C, Uehara, M. Improvement of Multi-Purpose Travel Route Recommendation System Based on Genetic Algorithm. In: 7th International Symposium on Computing and Networking Workshops (CANDARW); 2019 Nov 26-29; Nagasaki, Japan. IEEE; 2019, p. 305-308, doi:10.1109/CANDARW.2019.00060
- [5] Orama JA, Borràs J, Moreno A. Combining Cluster-Based Profiling Based on Social Media Features and Association Rule Mining for Personalised Recommendations of Touristic Activities. *Appl Sci*. 2021 Jul; 11(14):6512, doi:10.3390/app11146512.

On Conjunctive and Disjunctive Rational Bivariate Aggregation Functions of Degrees (2,1)

Isabel AGUILÓ^{a,b,1}, Sebastia MASSANET^{a,b} and Juan Vicente RIERA^{a,b}

^aSCOPIA research group, Dept. of Mathematics and Computer Science, University of the Balearic Islands, 07122 Palma, Spain

^bHealth Research Institute of the Balearic Islands (IdISBa), 07010 Palma, Spain

Abstract. In the last decades, many families of aggregation functions have been presented playing a fundamental role in many research fields such as decision making, fuzzy mathematical morphology, etc. For this reason, it is necessary to study different types of operators to be potentially used in a concrete application as well as the properties they can satisfy. In this paper, conjunctive and disjunctive rational bivariate aggregation functions of degree two in the numerator and degree one in the denominator are studied. In particular, a characterization of conjunctive and disjunctive rational aggregation functions of degrees (2,1) is presented. Moreover, the symmetry property of these operators are investigated.

Keywords. Aggregation function, Rational function, Symmetry

1. Introduction

Aggregation functions have been intensively investigated for about two decades, see [2, 3,4,6], playing a fundamental role in social and scientific sciences. There exist a great quantity of different aggregation functions satisfying different additional properties, each one having its particular importance and interest in different fields of application. In this sense, in this work we focus on two general classes: the class of conjunctive aggregation function (those that take values below the minimum that include well-known t-norms [5] or copulas [7]) and the class of disjunctive aggregation functions (those that take values over the maximum, that include, for instance, t-conorms or co-copulas [5]). All of them are used in meaningful applications in fuzzy logic and approximate reasoning, image processing, probability, statistics and economy [2,3].

In a previous work [1], rational binary aggregation functions, that is, those whose expression is given by the quotient of two bivariate polynomial functions were investigated. In particular, rational binary aggregation functions of degree (1,1) (both bivariate polynomials have degree 1) were characterized. Also, concrete characterizations of those that are symmetric and idempotent were also studied. Following this investigation line,

¹Corresponding Author: Isabel Aguiló, Dept. of Mathematics and Computer Science, University of the Balearic Islands, Ctra. de Valldemossa, Km.7.5, 07122 Palma, Spain; E-mail: isabel.aguilo@uib.es.

in this short paper we will study rational bivariate aggregation functions of degree two in the numerator and degree one in the denominator, with the focus on their characterization. These results are novel with respect to the ones presented in [1].

The paper is structured as follows. After recalling the basic definitions, in Section 3 conjunctive rational binary aggregation functions of degree $(2, 1)$ and those that are symmetric will be characterized. Section 4 is devoted to characterize disjunctive rational binary aggregation functions of degree $(2, 1)$ and those that are symmetric. The paper ends with some conclusions and future work we want to investigate.

2. Preliminaries

Let us recall some concepts and results that will be used throughout this paper. First, we give the definition of a binary aggregation function.

Definition 1 ([2,3]). *A binary aggregation function $f : [0, 1]^2 \rightarrow [0, 1]$ is a binary mapping that satisfies the following properties:*

- i) $f(0, 0) = 0$ and $f(1, 1) = 1$.
- ii) f is increasing in each variable.

A binary aggregation function f is a conjunction when $f(1, 0) = f(0, 1) = 0$ and a disjunction when $f(1, 0) = f(0, 1) = 1$. Two additional properties of binary aggregation functions which will be used in this work are the *symmetry*, $f(x, y) = f(y, x)$ for all $x, y \in [0, 1]$ and the *idempotency*, $f(x, x) = x$ for all $x \in [0, 1]$.

Now, we recall the definition of a rational aggregation function, introduced in [1].

Definition 2 (Definition 3 in [1]). *Consider $n, m \in \mathbb{N}$. A binary operator $R : [0, 1]^2 \rightarrow [0, 1]$ is called a rational aggregation function of degree (n, m) if it is an aggregation function and its expression is given by*

$$R(x, y) = \frac{p(x, y)}{q(x, y)} = \frac{\sum_{\substack{0 \leq i, j \leq n \\ i+j \leq n}} a_{ij} x^i y^j}{\sum_{\substack{0 \leq s, t \leq m \\ s+t \leq m}} b_{st} x^s y^t} \quad (1)$$

for all $x, y \in [0, 1]$ where

- (i) $a_{ij} \in \mathbb{R}$ for all $0 \leq i, j \leq n$ and $i + j \leq n$ and there exists at least one a_{ij} with $0 \leq i, j \leq n$ and $i + j = n$ such that $a_{ij} = 1$,
- (ii) $b_{st} \in \mathbb{R}$ for all $0 \leq s, t \leq m$ and $s + t \leq m$ and there exist some $0 \leq s, t \leq m$ with $s + t = m$ such that $b_{st} \neq 0$,
- (iii) the polynomials $p(x, y)$ and $q(x, y)$ have no factors in common,
- (iv) $q(x, y) \neq 0$ for all $x, y \in [0, 1]$.

3. Conjunctive Rational Binary Aggregation Functions of degree $(2, 1)$

In this section we will characterize the conjunctive binary rational aggregation functions of degree $(2, 1)$. Thus, the expression of the rational function $R(x, y)$ can be written as

$$R(x,y) = \frac{a_{20}x^2 + a_{02}y^2 + a_{11}xy + a_{10}x + a_{01}y + a_{00}}{b_{00} + b_{10}x + b_{01}y}$$

for all $x, y \in [0, 1]$ where each $a_{ij}, b_{ij} \in \mathbb{R}$. Based on this expression, next result provides a characterization of conjunctive rational aggregation functions of degree (2, 1).

Theorem 1. *A binary operator $R : [0, 1]^2 \rightarrow [0, 1]$ is a conjunctive rational binary aggregation function of degree (2, 1) if, and only if, R is given by*

$$R(x,y) = \frac{(b_{00} + b_{10} + b_{01})xy}{b_{00} + b_{10}x + b_{01}y} \quad (2)$$

for all $x, y \in [0, 1]$ where the coefficients satisfy $b_{00} > 0$, $b_{00} + b_{01} > 0$, $b_{00} + b_{10} > 0$, $b_{00} + b_{10} + b_{01} > 0$ and each $b_{ij} \in \mathbb{R}$.

The following result characterizes the symmetric conjunctive rational binary aggregation functions of degree (2, 1).

Proposition 1. *A binary operator $R : [0, 1]^2 \rightarrow [0, 1]$ is a symmetric conjunctive rational binary aggregation function of degree (2, 1) if, and only if, R is given by*

$$R(x,y) = \frac{(2b_{10} + b_{00})xy}{b_{10}(x+y) + b_{00}}$$

for all $x, y \in [0, 1]$ where $b_{00} > 0$, $b_{00} + b_{10} > 0$, $b_{00} + 2b_{10} > 0$ and each $b_{ij} \in \mathbb{R}$.

Proposition 2. *There are no idempotent conjunctive rational aggregation functions of degree (2, 1).*

4. Disjunctive Rational Binary Aggregation Functions of degree (2,1)

In this section we will characterize disjunctive binary rational aggregation functions of degree (2, 1).

Theorem 2. *A binary operator $R : [0, 1]^2 \rightarrow [0, 1]$ is a disjunctive rational binary aggregation function of degree (2, 1) if, and only if, R is given by*

$$R(x,y) = \frac{a_{11}xy + a_{10}x + a_{01}y}{(a_{11} + a_{10})x + (a_{11} + a_{01})y - a_{11}} \quad (3)$$

for all $x, y \in [0, 1]$ where the coefficients satisfy $a_{11} < 0$, $a_{10} > 0$, $a_{01} > 0$, $a_{10} + a_{01} + a_{11} > 0$, $a_{10} + a_{11} > 0$, $a_{01} + a_{11} > 0$ and each $a_{ij} \in \mathbb{R}$.

Next proposition characterizes the symmetric disjunctive rational binary aggregation functions of degree (2, 1).

Proposition 3. A binary operator $R : [0, 1]^2 \rightarrow [0, 1]$ is a symmetric disjunctive rational binary aggregation function of degree $(2, 1)$ if, and only if, R is given by

$$R(x, y) = \frac{a_{11}xy + a_{10}(x + y)}{(a_{11} + a_{10})(x + y) - a_{11}}$$

for all $x, y \in [0, 1]$ where the coefficients satisfy $a_{11} < 0$, $a_{10} > 0$, $2a_{10} + a_{11} > 0$, $a_{10} + a_{11} > 0$ and each $a_{ij} \in \mathbb{R}$.

Proposition 4. There are no idempotent disjunctive rational aggregation functions of degree $(2, 1)$.

The following result relates both families of aggregation functions.

Proposition 5. $R(x, y)$ is a disjunctive rational aggregation functions of degree $(2, 1)$ iff $1 - R(1 - x, 1 - y)$ is a conjunctive rational aggregation functions of degree $(2, 1)$.

5. Conclusions and future work

Following with the investigation on rational aggregation functions started in [1], in this work we have characterized all conjunctive and disjunctive rational binary aggregation functions of degree $(2, 1)$ and, in particular, we have also characterized those that are symmetric. As future work, first, we want to further explore different additional properties of these operators such as associativity or the existence of neutral or absorbing elements. Second, we want to analyze their performance in edge detection through fuzzy morphological operators. Finally, rational binary aggregation functions of higher degrees could be worthy to study although the complexity of the results increases drastically.

Acknowledgment

This paper is part of the R+D+i Project PID2020-113870GB-I00- “Desarrollo de herramientas de Soft Computing para la Ayuda al Diagnóstico Clínico y a la Gestión de Emergencias (HESOCODICE)”, funded by MCIN/AEI/10.13039/501100011033.

References

- [1] I. Aguiló, S. Massanet and J.V. Riera. On rational aggregation functions, accepted in the Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2022), 2022.
- [2] G. Beliakov, A. Pradera and T. Calvo. Aggregation Functions: A Guide for Practitioners, volume 221 of Studies in Fuzziness and Soft Computing. Springer, 2007.
- [3] T. Calvo, G. Mayor and R. Mesiar. Aggregation Operators: New Trends and Applications. Studies in Fuzziness and Soft Computing. Physica-Verlag HD, 2002
- [4] J.C. Fodor, On rational uninorms, Proceedings of the First Slovakian–Hungarian Joint Symposium on Applied Machine Intelligence, Herlany, Slovakia (2003), 139– 147.
- [5] E.P. Klement, R. Mesiar, E. Pap. Triangular norms. Kluwer Academic Publishers, Dordrecht, 2000.
- [6] M. Mas, S. Massanet, D. Ruiz-Aguilera, J. Torrens, “A survey on the existing classes of uninorms,” Journal of Intelligent and Fuzzy Systems, 29, 1021–1037, 2015.
- [7] R.B. Nelsen. An introduction to copulas. Springer, New York, EUA, 2006.

Approximate and Optimal Solutions for the Bipartite Polarization Problem

Teresa ALSINET, Josep ARGELICH¹, Ramón BÉJAR and Santi MARTÍNEZ

INSPIRES Research Center – University of Lleida

Jaume II, 69 – 25001 Lleida, SPAIN

Abstract. In a recent work we introduced a problem about finding the highest polarized bipartition on a weighted and labeled graph that represents a debate developed through some social network, where nodes represent user's opinions and edges agreement or disagreement between users. Finding this target bipartition is an optimization problem that can be seen as a generalization of the maxcut problem, so we first introduced a basic local search algorithm to find approximate solutions of the problem. In this paper we go one step further, and we present an exact algorithm for finding the optimal solution, based on an integer programming formulation, and compare the performance of a new variant of our local search algorithm with the exact algorithm. Our results show that at least on real instances of the problem, obtained from Reddit debates, the approximate solutions obtained are almost always identical to the optimal solutions.

Keywords. Social Networks, Polarization, Combinatorial Optimization.

1. Introduction

The emergence of polarization in discussions on social networks, and the responsibility of companies in this problem, is a topic that is causing a significant interest among society. For example, Facebook has launched some initiatives to try to mitigate the factors that may be helping the spread of divisive content [5], even if this kind of content may be the one that produces the maximum attention of their users, so being also the one producing maximum economic benefit.

Because each social network company can have its own personal interest regarding when to control this kind of behaviour, one fundamental aspect is to define more transparent ways to monitor such possible non-desirable behaviours so that we can decide to act only in situations where we can deduce that polarization is taking place, and to a certain level of severity, because there is some objective value we can measure for this. So, in a previous work, we defined one such measure

¹Correspondence to: J. Argelich. INSPIRES Research Centre, University of Lleida. C/Jaume II, 69. Lleida, Spain. Tel.: +34 973702734; E-mail: josep.argelich@udl.cat.

and presented an initial approximate algorithm to compute this measure from a discussion in the Reddit social network [1].

In this paper, we present an exact algorithm for computing this measure and a new variant of the approximate algorithm. The exact algorithm is based on an integer programming formulation, inspired by existing good formulations for the maxcut problem, as our problem can be seen as a generalization of the maxcut problem. Our preliminary experimental results over a set of Reddit debates show that the solutions obtained with the approximate algorithms are almost identical to the optimal solutions found by the exact algorithm, although computed with much less time.

The structure of the rest of the paper is as follows. In Section 2, we present both the representation model of Reddit debates and the measure to quantify the polarization in a debate, studied and developed in [1]. In Section 3, we introduce an exact algorithm for finding the optimal solution, based on an integer programming formulation. Finally, in Section 4, we perform an empirical evaluation to compare the performance of our solving approaches when computing the bipartition of two different sets of Reddit discussions.

2. Problem Definition

Following [1], a *Reddit debate* Γ on a root comment r is a non-empty set of Reddit comments, that were originated as successive answers to the root comment r that contains a link to some news. To represent debates on Reddit, we use a two-sided debate tree model, where nodes are labelled with a binary value that denotes whether the comment is in agreement (1) or in disagreement (-1) with the root comment.¹

Definition 1 (Two-Sided Debate Tree) *Let Γ be a Reddit debate on a root comment r . A Two-Sided Debate Tree (**SDebT**) for Γ is a tuple $\mathcal{T}_S = \langle C, r, E, W, S \rangle$ defined as follows:*

- *For every comment c_i in Γ , there is a node c_i in C .*
- *Node $r \in C$ is the root node of \mathcal{T} .*
- *If a comment $c_1 \in C$ answers another comment $c_2 \in C$, there is a directed edge (c_1, c_2) in E .*
- *W is a labelling function of answers (edges) $W : E \rightarrow [-2, 2]$, where the value assigned to an edge $(c_1, c_2) \in E$ denotes the sentiment of the answer c_1 with respect to c_2 , from highly negative (-2) to highly positive (2).*
- *S is a labelling function of comments (nodes) $S : C \rightarrow \{-1, 1\}$, where the value assigned to a node $c_i \in C$ denotes whether the comment c_i is in agreement (1) or in disagreement (-1) with the root comment r and it is defined as follows:*
 - $S(r) = 1$ and

¹Note that this definition can be applied to other similar social networks.

- For all node $c_1 \neq r$ in C , $S(c_1) = 1$ if for some node $c_2 \in C$, $(c_1, c_2) \in E$ and either $S(c_2) = 1$ and $W(c_1, c_2) > 0$, or $S(c_2) = -1$ and $W(c_1, c_2) \leq 0$; otherwise, $S(c_1) = -1$.

Only the nodes and edges obtained by applying this process belong to C and E , respectively.

Given a Reddit debate on a (root) comment, we make its corresponding SDebT using the Python Reddit API Wrapper (PRAW, available at <https://github.com/praw-dev/praw>) to download its set of comments, and then we evaluate the sentiment for each edge $(c_1, c_2) \in E$ using the sentiment analysis software of [4] using the text of the comment c_1 .

Our goal in this work is to introduce and investigate a suitable algorithm that allows us to analyse and study the polarization of users in debates. To this end, we group comments of a debate by user, and we consider that the relationship between two users is defined from the agreement and disagreement relationships between the individual comments of these two users.

Next, we present our formalization of a User Debate Graph based on a Two-Sided Debate Tree, where now we aggregate all the comments of a same user into a single node that represents the user's opinion.

Definition 2 (User Debate Graph) Let Γ be a Reddit debate on a root comment r with users' identifiers $U = \{u_1, \dots, u_m\}$ and let $\mathcal{T}_S = \langle C, r, E, W, S \rangle$ be a SDebT for Γ . A User Debate Graph (UDebG) for \mathcal{T}_S is a tuple $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$, where:

- \mathcal{C} is the set of nodes of \mathcal{G} defined as the set of users' opinions $\{C_1, \dots, C_m\}$; i.e., $\mathcal{C} = \{C_1, \dots, C_m\}$ with $C_i = \{c \in \Gamma \mid c \neq r \text{ and } \text{user}(c) = u_i\}$, for all users $u_i \in U$.
- $\mathcal{E} \subseteq \mathcal{C} \times \mathcal{C}$ is the set of edges of \mathcal{G} defined as the set of interactions between different users in the debate; i.e., there is an edge $(C_i, C_j) \in \mathcal{E}$, with $C_i, C_j \in \mathcal{C}$ and $i \neq j$, if and only if for some $(c_1, c_2) \in E$ we have that $c_1 \in C_i$ and $c_2 \in C_j$.
- \mathcal{S} is an opinion weighting scheme for \mathcal{C} that expresses the side of users in the debate based on the side of their comments. We define \mathcal{S} as the mapping $\mathcal{S} : \mathcal{C} \rightarrow [-1, 1]$ that assigns to every node $C_i \in \mathcal{C}$ the value

$$\mathcal{S}(C_i) = \frac{\sum_{c \in C_i} S(c)}{|C_i|}$$

in the real interval $[-1, 1]$ that expresses the side of the user u_i with respect to the root comment, from strictly disagreement (-1) to strictly agreement (1), going through undecided opinions (0).

- \mathcal{W} is an interaction weighting scheme for \mathcal{E} that expresses both the ratio of positive interactions between the users' opinions and the overall sentiment between users by combining the individual sentiment values assigned to the responses between their comments.

We define \mathcal{W} as the mapping $\mathcal{W} : \mathcal{E} \rightarrow ([0, 1] \times [-2, 2])$ that assigns to every edge $(C_i, C_j) \in \mathcal{E}$ the pair of values $(p, w) \in ([0, 1] \times [-2, 2])$ defined as follows:

$$p = \frac{|(c_1, c_2) \in E \cap (C_i \times C_j) \text{ with } W(c_1, c_2) > 0|}{|(c_1, c_2) \in E \cap (C_i \times C_j)|} \quad \text{and}$$

$$w = \sum_{\{(c_1, c_2) \in E \cap (C_i \times C_j)\}} W(c_1, c_2) / |\{(c_1, c_2) \in E \cap (C_i \times C_j)\}|$$

where p expresses the ratio of positive answers from the user u_i to the user u_j in the debate, and w expresses the overall sentiment of the user u_i regarding the comments of the user u_j , from highly negative (-2) to highly positive (2).

Only the nodes and edges obtained by applying this process belong to \mathcal{C} and \mathcal{E} , respectively.

Given a User Debate Graph $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$, in [1] we introduced a model to measure the level of polarization in the debate between its users. We identified two characteristics that a polarization measure should capture. First, a polarized debate should contain a bipartition of \mathcal{C} into two sets (L, R) such that the set L contains mainly users in disagreement, the set R contains mainly users in agreement, and both sets should be similar in size. The second ingredient is the sentiment between users of L and R . A polarized discussion should contain most of the negative interactions between users of L and users of R , whereas the positive interactions, if any, should be mainly within the users of L and within the users of R .

To capture these two characteristics with a single value, we defined two different measures and their combination in a final one, referred to as *the bipartite polarization level*.

Definition 3 (Bipartite Polarization) Given a User Debate Graph $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$ and a bipartition (L, R) of \mathcal{C} , we define:

- The level of consistency and balance of (L, R) is a real value in $[0, 0.25]$ defined as follows:

$$\text{SC}(L, R, \mathcal{G}) = \text{LC}(L, \mathcal{G}) \cdot \text{RC}(R, \mathcal{G})$$

with:

$$\text{LC}(L, \mathcal{G}) = \frac{\sum_{\substack{C_i \in L, \\ \mathcal{S}(C_i) \leq 0}} -\mathcal{S}(C_i)}{|\mathcal{C}|}$$

and

$$\text{RC}(R, \mathcal{G}) = \frac{\sum_{\substack{C_i \in R, \\ \mathcal{S}(C_i) > 0}} \mathcal{S}(C_i)}{|\mathcal{C}|}.$$

- The sentiment of the interactions between users of different sides is a real value in $[0, 4]$ defined as follows:

$$SWeight(L, R, \mathcal{G}) = \frac{\sum_{\substack{(i,j) \in \mathcal{E} \cap \\ ((L \times R) \cup (R \times L))}} -c(p(i,j)) \cdot w(i,j)}{|\mathcal{E}|} + 2,$$

with

$$c(p(i,j)) = 2(p(i,j) - 0.5)^2 + 1/2,$$

and where $p(i,j)$ and $w(i,j)$ denote the values of p and w , respectively, in $\mathcal{W}(i,j) = (p, w)$.

Then, the Bipartite Polarization of \mathcal{G} on a bipartition (L, R) is the value in the real interval $[0, 1]$ defined as follows:

$$BipPol(L, R, \mathcal{G}) = SC(L, R, \mathcal{G}) \cdot SWeight(L, R, \mathcal{G}).$$

Finally, the Bipartite Polarization of \mathcal{G} is the maximum value of $BipPol(L, R, \mathcal{G})$ among all the possible bipartitions (L, R) .

3. Algorithms

We can find the Bipartite Polarization of a User Debate Graph $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$ by solving the following integer nonlinear programming formulation (MINLP) of it, where each node C_i from \mathcal{C} is associated with an integer variable x_i , such that $x_i = -1$ represents that C_i is in the L partition and $x_i = +1$ that C_i is in R :

$$\begin{aligned} \max_x & \left(\frac{1}{|\mathcal{C}|^2} \sum_{\substack{(x_i, x_j) \text{ with} \\ S(C_i) \leq 0, S(C_j) > 0}} -S(C_i)S(C_j)(1 - x_i)(1 + x_j)/4.0 \right) * \\ & \left(2 + \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} -c(p(i,j)) \cdot w(i,j) * (1 - x_i * x_j)/2.0 \right) \\ & \text{subject to: } x_i^2 = 1 \quad \forall C_i \in \mathcal{C} \end{aligned}$$

Observe that the first sumatory in the objective function represents the term $SC(L, R, \mathcal{G})$ and the second one the term $SWeight(L, R, \mathcal{G})$.

Then, we use the branch-and-bound solver [6] of the SCIP Optimization suite (version 8.0) [3] to optimally solve problem instances with this MINLP formulation.

We can also use approximate algorithms based on local search, like the one we introduced in our previous work [1]. In that algorithm, the search ends as soon as the algorithm reaches a local maximum.

Now in this work, we present a slight variant where the algorithm performs diversification steps (non-improving steps) to try to escape from local maximum and be able to find better solutions later on. The pseudocode of this new variant, that is inspired by a local search for the maxcut problem [2], is shown on Algorithm 1. The search starts with some initial pseudo-random partition [1] (line 1), and then it initiates the search for a local maximum of the objective value, selecting at every step a node v that represents the steepest ascent hill climbing step, if such a move exists (line 6). That is, a node v that when swapped between L and R is the one that improves more the objective value. Each time the algorithm reaches a local maximum (line 9) it updates the best solution found so far (if necessary) and then it selects randomly some vertices to be swapped (line 12), where the probability that a node is selected is controlled with the parameter ω . After the diversification steps, the whole process is repeated, up to $max_restarts$ times.

Algorithm 1 Finding a local-optimal solution for the bipartite polarization of $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$, using diversification.

Input: $\mathcal{G} = \langle \mathcal{C}, \mathcal{E}, \mathcal{S}, \mathcal{W} \rangle$

Output: a bipartition (L, R) of \mathcal{G} with high bipartite polarization

```

1:  $(L, R) := getInitBPart(\mathcal{G})$                                  $\triangleright$  Get Initial bipartition
2:  $bestBipPol := getBipPol(L, R)$   $\triangleright$  Save as best bipartite polarization value...
3:  $(L', R') := (L, R)$                                           $\triangleright$  and save corresponding bipartition
4: for all  $max\_restarts$  do
5:   do
6:     if  $\exists v \in SAHC(L, R, \mathcal{G})$  then            $\triangleright$  Steepest Ascent Hill Climbing
7:        $swapNode(L, R, v)$                           $\triangleright$  Change the set of node  $v$ 
8:     while  $(\exists v \in SAHC(L, R, \mathcal{G})$  and not  $max\_steps)$ 
9:     if  $getBipPol(L, R) > bestBipPol$  then           $\triangleright$  Save new best solution
10:     $bestBipPol := getBipPol(L, R)$ 
11:     $(L', R') := (L, R)$ 
12:     $diversify(L, R, \omega)$                        $\triangleright \omega$  is the probability of diversification for a node
13: return  $(L', R')$ 

```

4. Experimental Results

We present here an small empirical evaluation of the performance of our three solving approaches (exact algorithm, previous local search algorithm [1] and improved local search algorithm) when computing the bipartite polarization of two different sets of Reddit discussions. We compare both the running time and how close are the approximate solutions obtained by the local search algorithms compared with the optimal solution. The results are shown on Table 1, where for

	Stats	$ \mathcal{C} $	SCIP			LS1		LS2	
			time	nodes	BipPol	time	ratio	time	ratio
DS1	min	17	0.11	0	0.166071	0.01	0.934859	0.01	0.999998
	median	42	1.19	1	0.478105	0.01	1	0.06	1
	mean	41	6.48	1.75	0.470922	0.01	0.998220	0.08	0.999999
	max	86	104.4	38	0.610654	0.02	1	0.35	1
DS2	min	25	0.2	0	0.443185	0.01	0.999592	0.02	1
	median	50	3.94	1	0.540176	0.01	1	0.11	1
	mean	53	24.59	1.64	0.543161	0.01	0.999988	0.15	1
	max	102	337.51	35	0.674507	0.04	1	0.65	1

Table 1. Experimental results with algorithms for computing bipartite polarization.

the exact algorithm (SCIP) we show statistics (min, median, mean and max) for its solving time over the instances of the first dataset (DS1) and over the second dataset (DS2), as well as the same statistics but for the number of nodes of the branch-and-bound search tree and the bipartite polarization value of the instances (BipPol). We also show the same statistics for the number of vertices (users of the user debate graph) in the two different data sets, so we can observe that the second dataset contains slightly bigger instances than the first one. We also show the results for the two local search approaches, the previous one (LS1) and the new presented in this paper (LS2). The values shown are their execution times and the ratio of the solution obtained by the local search algorithm to the optimal solution found by SCIP.

As the results indicate, the local search algorithms almost always find the optimal solution, but with a much smaller running time. However, it is interesting to note that the solving time, and number of nodes of the search tree, is in general quite small for the SCIP solver, as the median time is in both datasets around 2 seconds, and the median number of nodes is one. The fact that SCIP is able to solve most of the instances with no branching at all, is due to the fact that in those instances it is able to solve them only during the “presolving” phase, where it uses different simplification techniques, although the mean and max values show that there are instances where presolving is not enough. It is not clear how these results will be for bigger instances, but for the instances tested here it is clear that the local search approaches seem to be enough to solve the instances. Regarding the differences between LS1 and LS2, we observe that already LS1 is able to find almost always the optimal solution, but LS2 seems to obtain a better ratio (almost always equal to 1) in all the cases.

5. Conclusions

In this paper we present several algorithms to solve the Bipartite Polarization Problem, that can be seen as a generalization of the maxcut problem. To this end, we first introduce two variants of a basic local search algorithm to find approximate solutions. Next, we also develop a complete algorithm based on the integer nonlinear programming formulation. Both approaches show very good performance, being the incomplete one the fastest and the complete one the more

accurate. Also, as we can see in the results, the incomplete approach is pretty accurate as the solutions are always very close to the optimal solution.

Further experimental results will be needed to understand their scaling behavior as the size of the instances increases. As further work, we also plan to explore other solving techniques, like the ones based on Goemans-Williamson's Semidefinite Positive relaxation, and study what features make the instances easier to be solved with the local search approach.

Acknowledgments This work was partially funded by Spanish Project PID2019-111544GB-C22 (MINECO / FEDER), by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreements 723596, 768824, 764025 and 814945, and by 2017 SGR 1537.

References

- [1] T. Alsinet, J. Argelich, R. Béjar, and S. Martínez. Measuring polarization in online debates. *Applied Sciences*, 11(24), 2021.
- [2] U. Benlic and J. Hao. Breakout local search for the max-cutproblem. *Eng. Appl. Artif. Intell.*, 26(3):1162–1173, 2013.
- [3] K. Bestuzheva, M. Besançon, W.-K. Chen, A. Chmiela, T. Donkiewicz, J. van Doornmalen, L. Eifler, O. Gaul, G. Gamrath, A. Gleixner, L. Gottwald, C. Graczyk, K. Halbig, A. Hoen, C. Hojny, R. van der Hulst, T. Koch, M. Lübbecke, S. J. Maher, F. Matter, E. Mühmer, B. Müller, M. E. Pfetsch, D. Rehfeldt, S. Schlein, F. Schlösser, F. Serrano, Y. Shinano, B. Sofranac, M. Turner, S. Vigerske, F. Wegscheider, P. Wellner, D. Weninger, and J. Witzig. The scip optimization suite 8.0. Technical Report 21-41, ZIB, Takustr. 7, 14195 Berlin, 2021.
- [4] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.
- [5] G. Rosen. Facebook: Investments to fight polarization. <https://about.fb.com/news/2020/05/investments-to-fight-polarization/>, 2020. date: 2020-05-27.
- [6] S. Vigerske and A. Gleixner. Scip: global optimization of mixed-integer nonlinear programs in a branch-and-cut framework. *Optimization Methods and Software*, 33(3):563–593, 2018.

Yet Another (Fake) Proof of P=NP

Carlos ANSÓTEGUI^a and Jordi LEVY^{b,1}

^aUniversitat de Lleida, Spain.

^bIIIA, CSIC, Spain.

Abstract. Obviously, we do not prove $P = NP$ in this article. In fact, the title only refers to the first part, where the proof that we present contains an error that, to make reading more attractive, is only revealed in the second part.

In the second part, we describe how the reduction of SAT to Max2XOR and the proof system presented in the first part –although they do not solve one of the Millennium Prize Problems– may trigger new complementary ways of solving the SAT problem.

Keywords. Proof systems, SAT, MaxSAT, MaxCUT

1. Introduction

We start this fake proof by introducing the Cook and Reckhow program [1], as a way to resolve the **P** versus **NP** problem. The program is based on the following observation. The complexity class **P** is closed under complement, therefore, if $P = NP$, then **NP** is also closed under complement. In other words,

$$NP \neq CoNP \text{ implies } P \neq NP \quad (1)$$

The class **P** does not need any introduction. The class **NP** is the set of decision problems that can be solved in polynomial time by a non-deterministic Turing Machine. Equivalently, **NP** can be defined as the class of decision problems P for which $x \in P$ has *certifications* or *proofs* verifiable in polynomial time by a deterministic Turing machine. The classical **NP-hard** problem is SAT [2,3], defined as the set of propositional formulas in conjunctive normal form that are satisfiable. In this case, a certification that a given formula is in SAT can be simply a truth assignment to the variables that satisfies all the clauses. Given a formula and a truth assignment, we can verify that the assignment certifies the satisfiability of the formula using a deterministic Turing Machine in polynomial time.

The class **CoNP** is the complement of **NP**. It can be defined as the class of decision problems P for which $x \notin P$ has certifications verifiable in polynomial time by a deterministic Turing machine. The classical **CoNP-hard** problem is the complement of SAT, i.e. TAUT.

In order to prove that $NP = CoNP$, since TAUT is **CoNP-hard**, we only need to prove that $x \in TAUT$ has certifications verifiable in polynomial time. Here, it makes

¹Corresponding Author: Jordi Levy, IIIA, CSIC, Campus de la UAB, 08193 Bellaterra, Spain; E-mail: levy@iiia.csic.es.

sense to call these certifications *proofs*. These proofs are defined in a very general way, as sequences of bits $s \in \{0, 1\}^+$. A *proof system* PS is defined as a polynomial-time algorithm that, for any formula f , we have $f \in TAUT$ if, and only if, there exists a proof s such that the algorithm PS accepts (f, s) . We say that the proof system is *p-bounded* if, in addition, we can ensure the existence of a polynomially-bounded proof. In this case, PS accepts (f, s) in polynomial time on $|f|$. Therefore,

$$\text{there is a p-bounded proof system for TAUT if, and only if, } \mathbf{NP} = \mathbf{CoNP}. \quad (2)$$

It is important to remark that we do not care about how difficult is to find the proof. We only have to ensure that there exists a *short* (i.e. p-bounded) proof. A completely different question is the practical use of these proof systems, in a SAT solver, for instance. Thus, we observe that most SAT solvers are based on the *resolution* proof system, although we know that it is one of the weakest proof systems, in the sense that it has long proofs for tautologies that have short proofs in other proof systems.

2. Reducing SAT to Max2XOR

XOR clauses are similar to SAT clauses, using exclusive OR instead of the traditional OR. They can be written in the form of parity constraints or clauses

$$x_1 \oplus \dots \oplus x_n = k,$$

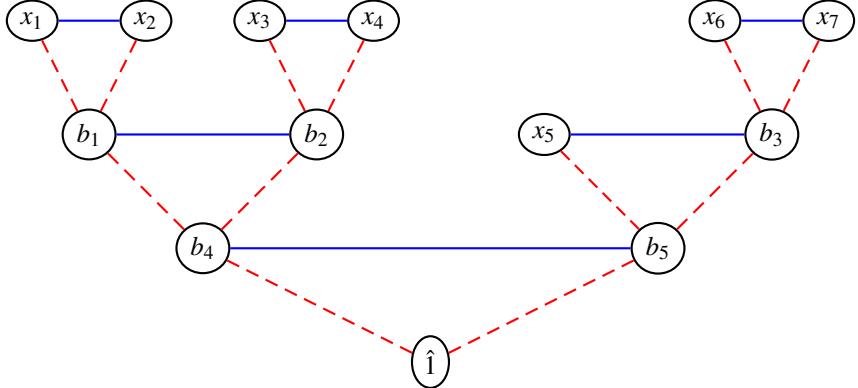
where x_i 's are Boolean variables (that may take value 0 or 1), \oplus is the sum-modulo-two operator, and k is either the constant 0 or 1. The particular case where $n = 0$ and $k = 1$ is an unsatisfiable clause, similar to the empty clause in SAT, that we will represent as \square . The case where $n = 0$ and $k = 0$ is always satisfied (a tautological clause) and is removed from the formula. Repeated variables are also removed from clauses, thus $x \oplus x \oplus A = k$ is simplified to $A = k$.

Here we will consider parity constraints with at most two variables. The satisfiability of a set of these constraints, called 2XOR –or even the more general XOR problem– is in \mathbf{P} .

Max2XOR is the optimization version of this problem, where we try to find an assignment that maximizes the number of satisfied constraints. The decision version of this problem, where given a set of parity constraints and an integer C , we decide if there is an assignment that satisfies at least C constraints, is \mathbf{NP} -hard. This can be proved by reducing SAT to Max2XOR using the gadget we describe in [4]. Below, we introduce this reduction through an example.²

Example 1. The clause $x_1 \vee \dots \vee x_7$ can be reduced to the following Max2XOR problem, where a dashed red line between x and y represents the clause $x \oplus y = 0$, or equivalently the equal-variable constraint $x = y$, and a solid blue line between x and y represents the clause $x \oplus y = 1$, or equivalently the non-equal-variable constraint $x \neq y$. The special node labeled as $\hat{1}$ is used to represent unary clauses as edges: $x \oplus \hat{1} = 1$ is equivalent to $x = 0$, and $x \oplus \hat{1} = 0$ is equivalent to $x = 1$. The variables b_1, \dots, b_5 are fresh variables not occurring elsewhere.

²We refer to [4] for a formal definition and proof of correctness. In this original publication, XOR clauses are weighted, and all weights are multiples of $1/2$. Here we avoid the use of weights.



Any similar set of XOR clauses, representing this kind of tree with the variables x_1, \dots, x_7 in the leaves, would also work.

If the original clause contains negated literals, we proceed in the same way, transforming the resulting XOR clauses by removing negations, using: $\neg x \oplus A = k$ is equivalent to $x \oplus A = 1 - k$.

Proposition 2 ([4]). *The Translation of a SAT clause C with $n \geq 2$ literals results into a multiset $T(C)$ of $3(n - 1)$ XOR clauses.*

For any SAT clause C of size $n \geq 2$: any assignment I satisfying C , can be extended (assigning the auxiliary variables b 's) to an assignment satisfying $2(n - 1)$ XOR clauses of $T(C)$, and any assignment I falsifying C can only be extended to satisfy at most $2(n - 2)$ of the XOR clauses.

Any SAT problem P with clause sizes $C_i \geq 2$, for $i = 1, \dots, m$, can be reduced to a Max2XOR problem $T(P)$ with $3\sum_{i=1}^m (C_i - 1)$ XOR clauses, such that

P is satisfiable if, and only if, the maximum number of satisfiable clauses in $T(P)$ is at least $2\sum_{i=1}^m (C_i - 1)$; or equivalently,

P is unsatisfiable if, and only if, the minimum number of unsatisfied clauses in $T(P)$ is at least $2 + \sum_{i=1}^m (C_i - 1)$.

The previous proposition allows us to reduce SAT to (the decision version of) Max2XOR. In the next section we tackle the problem of defining a proof system for Max2XOR or, in general, for MaxXOR.

MaxCUT is the problem of, given a graph, finding a partition of the vertices into two sets such that the number of edges connecting two nodes of distinct partitions is maximized. It is easy to see that MaxCUT is the same as maximizing the satisfaction of constraints of the form $x \oplus y = 1$. We can reduce Max2XOR to MaxCUT as follows: We introduce a new variable, called $\hat{1}$, and replace every constraint $x = 0$ by $x \oplus \hat{1} = 1$, and $x = 1$ by $x \oplus \hat{1} = 0$. Then, we replace every constraint of the form $x \oplus y = 0$ by two constraints $x \oplus b = 1$ and $b \oplus y = 1$, where b is a fresh variable. In this way, we get a set of constraints of the form $x \oplus y = 1$ that, interpreted as edges, can be solved as a MaxCUT problem.

3. A Polynomial Proof System for MaxXOR

The equivalent to the resolution rule for XOR constraints, called *XOR resolution rule* or *Gaussian elimination*, is

$$\frac{\begin{array}{l} x \oplus A = k_1 \\ x \oplus B = k_2 \\ \hline A \oplus B = k_1 \oplus k_2 \end{array}}{} \quad (3)$$

When we apply this rule, we must remove repeated variables (we assume that there is at least one of such repeated variables, called x in the previous scheme). Thus, these are particular instances of the rule scheme:

$$\frac{\begin{array}{l} x \oplus y \oplus z = 1 \\ x \oplus y \oplus t = 1 \\ z \oplus t = 0 \end{array}}{\square} \quad \frac{\begin{array}{l} x \oplus y = 1 \\ x \oplus y = 0 \end{array}}{\square}$$

where, in addition to x , we also remove the repeated variable y . Based on this rule we can prove that deciding satisfiability of XOR constraints is in **P**.

In order to extend this rule to MaxXOR or Max2XOR, we can take inspiration from the extension of resolution to MaxSAT resolution [5,6]. The MaxSAT resolution rule is:

$$\frac{\begin{array}{l} x \vee A \\ \neg x \vee B \\ \hline A \vee B \end{array}}{} \quad \frac{\begin{array}{l} x \vee A \vee \neg B \\ \neg x \vee B \vee \neg A \end{array}}{}$$

where the rule deals with multisets of clauses, *replacing* premises by conclusions, and where, if $B = b_1 \vee \dots \vee b_n$, the meta-clause $x \vee A \vee \neg B$ stands for the set $\{x \vee A \vee \neg b_1, x \vee A \vee b_1 \vee \neg b_2, \dots, x \vee A \vee b_1 \vee \dots \vee b_{n-1} \vee \neg b_n\}$. This rule is sound and complete for MaxSAT, in the sense that, m is the minimum number of unsatisfiable clauses in a multiset Γ if, and only if, there exists a derivation $\Gamma \vdash \{\square, \dots, \square\} \cup \Delta$, where Δ is a multiset of satisfiable clauses and \square stands for the empty clause. The proof of the soundness of this rule is based on the fact that, for any assignment, the number of falsified premises is equal to the number of falsified conclusions. The proof of completeness is more complicated. Basically, we prove that we can always construct a regular refutation in the following way. We consider a fixed list of variables x_1, \dots, x_n . We only resolve the variable x_1 until all occurrences of this variable have the same sign, or all possible pairs are of the form $x_1 \vee A$ and $\neg x_1 \vee B$ with a variable y satisfying $y \in A$ and $\neg y \in B$. In this situation, the new clause $A \vee B$ is a tautology. Then, we proceed to resolve x_2 , and so on. In the end, we only have empty clauses and satisfiable clauses.

In the case of MaxXOR, simply consider the same XOR-resolution rule (3), but applied *replacing* premises by the conclusion. First, we have the following observation:

Lemma 3. *For any assignment, in the XOR-resolution rule, the number of falsified premises is always equal to or greater than the number of falsified conclusions.*

The number of falsified clauses only decreases when the assignment falsifies both premises.

Proof. An analysis of all the cases allows us to verify that, for any assignment, there are only three possibilities: 1) the assignment satisfies both premises and the conclusion, 2) the assignment satisfies one of the premises and falsifies the other premise and the conclusion, or 3) the assignment falsifies both premises and satisfies the conclusion. \square

This fact is enough to prove the soundness of a proof system based on this rule:

Lemma 4 (Soundness). *The XOR-resolution rule is sound for MaxXOR, i.e. if there exists a refutation of the form $\Gamma \vdash \{\square, \dots, \square\} \cup \Delta$, then the minimal number of unsatisfiable constraints in Γ is at least m .*

Proof. For any assignment I , by Lemma 3, the number of clauses falsified by I in Γ is bigger or equal to the number of clauses falsified by I in $\{\square, \dots, \square\} \cup \Delta$, that is at least m because empty clauses are always falsified. Therefore, m is a minimum of the number of clauses falsified in Γ for any assignment. \square

Completeness is more complicated. As for the MaxSAT resolution rule, to construct the proof $\Gamma \vdash \{\square, \dots, \square\} \cup \Delta$, we can fix an ordering of the variables x_1, \dots, x_n . Then, we apply XOR-resolution to remove all pairs of occurrences of x_1 , except those that generate tautologies.³ When finished, we continue with x_2 , and so on until we get the empty clauses. However, in this process, we have to ensure that, at least for one assignment, the number of unsatisfied clauses is preserved. Therefore, using this *guiding* assignment, we will avoid applications of the XOR-resolution when the two premises are falsified and the conclusion is satisfied. As we have seen in the proof of Lemma 3, this is the only situation where the number of unsatisfied clauses is not preserved.

Given an assignment I , we write $\Gamma \vdash_I \Delta$ when we can derive Δ from Γ using the MaxXOR inference rule (i.e. replacing premises by the conclusion) and, in all steps, the assignment satisfies at least one of the premises. This ensures that the number of clauses falsified by I in Γ is equal to the number of clauses falsified by I in Δ .

The following lemma states that, for any variable x , we can always choose an assignment –in fact, any of the optimal assignments will work– that allows us to derive $\Gamma \vdash_I \Delta$ where Δ only contains one kind of clause for x . Basically, this allows us to forget about x and, *eventually*, proceeding in the same way with the rest of the variables, removing all of them, obtaining at the end a set of empty clauses. Since the set of unsatisfied clauses is preserved, the cardinality of this multiset of empty clauses would be equal to the number of clauses falsified by I in the original problem, i.e. equal to the minimum number of unsatisfiable clauses.

Lemma 5 (Iteration). *For any XOR formula Γ and any variable x , there exists an assignment I and a derivation $\Gamma \vdash_I \Delta$, where all clauses in Δ that contain x are equal. In other words, $\Gamma \vdash_I C_x \cup \Delta'$, where Δ' does not contain x and $C_x = \{x \oplus A = k, \dots, x \oplus A = k\}$.*

Proof. Let I be any optimal assignment that maximizes the number of satisfied clauses. Let Γ_x be the subset of clauses that contain x . Since I is optimal, the number of clauses of Γ_x satisfied by I is bigger than the number of clauses of Γ_x falsified by I . If this were not true, we could consider another assignment I' defined as $I'(y) = I(y)$, if $x \neq y$, and $I'(x) = 1 - I(x)$ that satisfies more clauses than I . Then, we can partition $\Gamma_x = \bigcup_{i=1}^r A_i \cup$

³Notice that the only way to obtain a tautology is to resolve a clause with another identical clause, i.e. if we resolve $x \oplus A = k$ and $x \oplus A = k$. Therefore, we only stop when all clauses containing x_1 are equal.

$\bigcup_{j=1}^s B_j \cup C$, where A_i contain a couple of distinct clauses both satisfied by I , B_j contain a pair of clauses, one satisfied by I and the other falsified by I , and all clauses in C are equal. We can apply \vdash_I to every pair in A_i 's and in B_j 's, leaving C as the only clauses that contain x . \square

Notice that, since there is only one kind of clause containing x , these clauses $x \oplus A = k$ in C_x are always satisfiable by taking $I(x) = k - I(A)$.

Finally, we can prove that any XOR-resolution proof is polynomially bounded:

Lemma 6. *For any Γ , the length of any proof $\Gamma \vdash \Delta$ is linearly bounded on $|\Gamma|$.*

Proof. It is trivial, since any XOR-resolution step removes one clause from the multiset. \square

4. Where is the Bug?

The attentive reader can probably skip this section, or probably wants to think about the MaxXOR proof system a little bit more before we reveal the solution..., no? Then,...

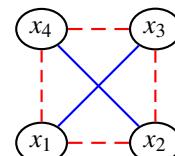
In fact, to the best of our knowledge, everything we have said so far is true! However, it is said in a way that might lead the reader to conclude that this proves $\mathbf{P} = \mathbf{NP}$. Actually, there are two *pitfalls*.

The first one is that, even if we were able to find a p-bounded proof system for TAUT, we cannot conclude $\mathbf{P} = \mathbf{NP}$. Notice that statement (2) is a double implication, but statement (1) only works in one direction. This means that we could prove $\mathbf{NP} = \mathbf{CoNP}$, and make all the arithmetic hierarchy collapse, but we could still have $\mathbf{P} \subsetneq \mathbf{NP} = \mathbf{CoNP}$. This possibility is not usually considered when discussing the \mathbf{P} versus \mathbf{NP} problem, but it is perfectly plausible. It imply that we could no longer try to use the Cook and Reckhow's approach to solve the \mathbf{P} versus \mathbf{NP} problem.

The second pitfall makes the proof system that we have defined for MaxXOR incomplete. This can be better seen through the following example.

Example 7. Consider the following MaxXOR problem:

$$\begin{array}{ll} x_1 \oplus x_2 = 0 & x_1 \oplus x_3 = 1 \\ x_2 \oplus x_3 = 0 & x_2 \oplus x_4 = 1 \\ x_3 \oplus x_4 = 0 & \\ x_4 \oplus x_1 = 0 & \end{array}$$

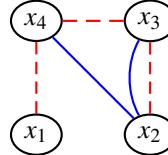


where, on the right, we represent the 2XOR clauses with red dashed lines to represent equal-variable constraints, and with blue lines to represent non-equal-variable constraints.

The minimum number of unsatisfiable clauses is 2 and there are several optimal assignments. One of them is $I(x_1) = I(x_2) = I(x_3) = I(x_4) = 1$. Using this optimal assignment, the four equal-variable constraints are satisfied, and the two non-equal-variable constraints are falsified. Assume that we want to start by removing variable x_1 . As shown in the proof of Lemma 5, among the clauses that contain x_1 , there are more satisfied than falsified by I , and we can arrange them in the partition $B_1 = \{x_1 \oplus x_2 = 0, x_1 \oplus x_3 = 1\}$

and $C = \{x_4 \oplus x_1 = 0\}$. Applying one XOR resolution step, we replace B_1 by $\{x_2 \oplus x_3 = 1\}$, leaving C frozen:

$$\begin{aligned} x_2 \oplus x_3 &= 1 \\ x_2 \oplus x_3 &= 0 \quad x_2 \oplus x_4 = 1 \\ x_3 \oplus x_4 &= 0 \\ x_4 \oplus x_1 &= 0 \end{aligned}$$



The problem is that I is no longer optimal for this new MaxXOR problem. This implies, for instance, that for some variables, the set of clauses containing it may contain more falsified clauses than satisfied. In our example, there are two clauses involving x_2 that are falsified by I , whereas only one is satisfied by I . Moreover, if we try to remove x_3 or x_4 , in both cases, we reach a situation where we cannot continue the application of XOR resolution steps. In fact, whatever we do, our proof system only can obtain one empty clause from our original MaxXOR instance.

The previous example shows that, although Lemma 5 is true and allows us to remove one variable, the optimal assignment is not optimal for the resulting MaxXOR problem. Therefore, we can not use the same lemma to iteratively remove another variable. In the previous section we have added an “*eventually*” to avoid the introduction of false statements and used the name “*iteration*” for Lemma 5 that is misleading, since it only applies to the first variable.

Taking into consideration the reduction of Max2XOR to MaxCUT, or from SAT to MaxCUT, it is easy to see that the existence of polynomial certifications or proofs for the MaxCUT problem would have important consequences and would also prove $\text{NP} = \text{CoNP}$.

5. Three Pigeons and Two Holes

The Pigeon-hole principle, noted PHP_n^{n+1} , is a family of tautologies that state that we can not place $n + 1$ pigeons in n holes with at most one pigeon in each hole. In the form of SAT clauses, it is encoded as:

$$\begin{aligned} x_i^1 \vee \dots \vee x_i^n &\quad \text{for } i = 1, \dots, n+1 \\ \neg x_i^k \vee \neg x_j^k &\quad \text{for } 1 \leq i < j \leq n+1 \text{ and } k = 1, \dots, n \end{aligned}$$

where x_i^j means that pigeon i is placed in hole j . This principle was used by Haken [7] to prove that some tautologies require super-polynomial proofs in the resolution proof system.

In this section, we show how the pigeon-hole principle PHP_2^3 , with 3 pigeons and 2 holes, may be proved with the XOR-resolution proof system. We use the SAT to Max2XOR reduction described in Section 2. The clauses $x_i^1 \vee x_i^2$, for $i = 1, 2, 3$, generate the constraints $\{x_1^1 = 1, x_2^1 = 1, x_1^1 \oplus x_2^1 = 1\}$. The clauses $\neg x_i^j \vee \neg x_{i'}^j$, for $i, i' = 1, 2, 3$ and $i < i'$, and $j = 1, 2$, generate the constraints $\{x_i^j = 0, x_{i'}^j = 0, x_i^j \oplus x_{i'}^j = 1\}$. All these constraints are represented in Figure 1. From them, only resolving the constraints with the same colors, we get 11 copies of \square . The Max2XOR problem comes from the translation of 9 binary clauses. Therefore, according to Proposition 2, we had to get $2 + 9(2 - 1) = 11$

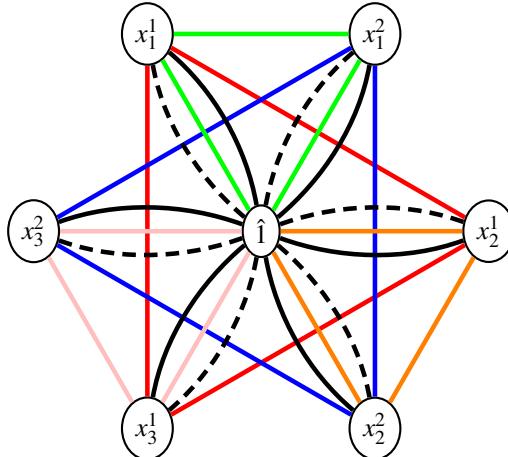


Figure 1. Graphical representation of the proof of PHP_2^3 . Solid lines between two nodes x and y represent $x \oplus y = 1$, and between x and 1 represents $x = 0$. Similarly, dashed lines represent $x \oplus y = 0$ or $x = 1$. Every pair of black lines with the same origin and target generates an empty clause (there are 6 pairs). For the rest 5 colors, each set of 3 lines of the same color that forms a triangle generates another empty clause.

copies of \square to prove the unsatisfiability of the original formula. This concludes the proof of PHP_2^3 .

6. Makes It Sense to Reduce SAT to Max2XOR?

A priori, the reduction of SAT to Max2XOR has some pros and cons. As for cons, since we are trying to maximize the number of satisfied clauses, instead of satisfying all of them, it is more difficult to make inferences. For instance, it is difficult to apply learning techniques, which have been so successful in modern SAT solvers. In the case of MaxSAT, we can use MaxSAT resolution (instead of classical resolution) to make inferences. In the case of MaxXOR, it is difficult to obtain a similar complete proof system. As an advantage, we can mention that the transformation produces *small* clauses. Anybody familiarized with SAT techniques would appreciate the generation of these small clauses, because they reduce drastically the number of possible satisfying assignments. In [8], we show how the reduction of SAT to MaxSAT and the use of MaxSAT resolution on the resulting formula, allows us to obtain polynomial proofs for the pigeon-hole principle that, as we mentioned, has super-polynomial proofs in resolution.

It is preferable to reduce SAT to Max2SAT, as in [8], or to Max2XOR, as in [4]? Again, a priori, there are advantages in both cases. In the case of Max2SAT, we have a complete proof system, that we do not have in the case of Max2XOR. However, in the case of Max2SAT, we have to consider that, even if all clauses are unary or binary, as a result of the application of the MaxSAT resolution rule, we can get bigger clauses, that have to be reduced again. In the case of MaxXOR, we have the advantage that the decision problem XOR is in **P**, whereas the decision problem for MaxSAT, i.e. SAT, is **NP**-complete. This means that, once we have decided which clauses we want to satisfy, it is easy to check if they are satisfiable.

The classical scheme of a SAT solver may be seen as a search engine that tries to find a satisfying assignment working together with an inference engine that tries to construct a proof of unsatisfiability. The great advantage of resolution as a proof system is that the construction of the proof is similar to the search of the assignment. Therefore, both engines work closely interleaved.

On the other side, so-called local search SAT solvers, only try to find a satisfying assignment, and can work forever if such an assignment does not exist. If we do not have a complete proof system for Max2XOR, it seems that we can only undertake to implement local-search solvers.

The best approximated algorithm for MaxCUT is based on a relaxation to Semidefinite Programming (SDP) [9]. In practice, although SDP is polynomial, it is still inefficient. One possibility is to reduce the number of dimensions used to represent the vectors associated with Boolean variables. It is known that, for a sufficient rank ($\sqrt{2n}$ instead of n), the solution of the approximating problem is still unique and the so-called *mixing method* converges to it [10]. The mixing method has been adapted to Max2SAT [11], and it can be adapted to the Max2XOR problem. We have started to explore this possibility, and the results obtained with this approximated algorithm are promising.

Once we have an approximated algorithm, we can apply several techniques on top of it. One possibility is to implement a *decimation* algorithm. It consists of, once we have a kind of probability for each variable to have a certain truth value, provided by the approximated algorithm, we can fix this value for one or a fraction of the variables for which this probability is higher, and we call again to the approximated algorithm. Another possibility is to use a branch-and-bound algorithm. In the case of Max2XOR, the approximated algorithm also returns a kind of probability for each clause to be satisfied. Therefore, in parallel, we can fix the values of variables and also force the satisfaction of clauses with higher probabilities.

Finally, we know that real-world SAT instances are highly modular [12,13]. The reduction defined in [4] allows us to group variables in any way. Therefore, we can analyze the modular structure of the original SAT instance, and get a Max2XOR problem where only variables closely related occur together in most of the clauses. This would contribute to increasing the modularity and thus make it easier to find a solution.

7. Conclusions and Further Work

We have used a fake proof of $P = NP$ as an excuse to introduce (we hope that in a funny way) a reduction of SAT to Max2XOR. This has also served us to comment on the difficulties of defining a complete proof system for Max2XOR. To finish, we have discussed the possibilities of implementing a local-search algorithm for Max2XOR based on the mixing method. The preliminary results using these ideas are promising. The modular structure of the original SAT instance can be used in the reduction to obtain an even more modular Max2XOR problem.

References

- [1] Cook SA, Reckhow RA. The Relative Efficiency of Propositional Proof Systems. *J Symb Log*. 1979;44(1):36-50. DOI: [10.2307/2273702](https://doi.org/10.2307/2273702).

- [2] Cook SA. The Complexity of Theorem-Proving Procedures. In: Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, STOC'71. ACM; 1971. p. 151-8. DOI: [10.1145/800157.805047](https://doi.org/10.1145/800157.805047).
- [3] Levin LA. Universal Sequential Search Problems. Problems of Information Transmission. 1973;9(3).
- [4] Ansótegui C, Levy J. Reducing SAT to Max2XOR. ArXiv. 2022;2204.01774. DOI: [10.48550/ARXIV.2204.01774](https://doi.org/10.48550/ARXIV.2204.01774).
- [5] Bonet ML, Levy J, Manyà F. A Complete Calculus for Max-SAT. In: Proceedings of the 9th International Conference on Theory and Applications of Satisfiability Testing, SAT'06. vol. 4121 of Lecture Notes in Computer Science. Springer; 2006. p. 240-51. DOI: [10.1007/11814948_24](https://doi.org/10.1007/11814948_24).
- [6] Bonet ML, Levy J, Manyà F. Resolution for Max-SAT. Artif Intell. 2007;171(8-9):606-18. DOI: [10.1016/j.artint.2007.03.001](https://doi.org/10.1016/j.artint.2007.03.001).
- [7] Haken A. The intractability of resolution. Theoretical Computer Science. 1985;39:297-308. Third Conference on Foundations of Software Technology and Theoretical Computer Science. DOI: [10.1016/0304-3975\(85\)90144-6](https://doi.org/10.1016/0304-3975(85)90144-6).
- [8] Ansótegui C, Levy J. Reducing SAT to Max2SAT. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI'21; 2021. p. 1367-73. DOI: [10.24963/ijcai.2021/189](https://doi.org/10.24963/ijcai.2021/189).
- [9] Goemans MX, Williamson DP. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. J ACM. 1995 nov;42(6):11151145. DOI: [10.1145/227683.227684](https://doi.org/10.1145/227683.227684).
- [10] Wang PW, Chang WC, Kolter JZ. The Mixing method: low-rank coordinate descent for semidefinite programming with diagonal constraints. ArXiv. 2018;1706.00476.
- [11] Wang P, Kolter JZ. Low-Rank Semidefinite Programming for the MAX2SAT Problem. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI'19. AAAI Press; 2019. p. 1641-9. DOI: [10.1609/aaai.v33i01.33011641](https://doi.org/10.1609/aaai.v33i01.33011641).
- [12] Ansótegui C, Giráldez-Cru J, Levy J. The Community Structure of SAT Formulas. In: Proceedings of the 15th International Conference on Theory and Applications of Satisfiability Testing, SAT'12. vol. 7317 of Lecture Notes in Computer Science. Springer; 2012. p. 410-23. DOI: [10.1007/978-3-642-31612-8_31](https://doi.org/10.1007/978-3-642-31612-8_31).
- [13] Ansótegui C, Bonet ML, Giráldez-Cru J, Levy J, Simon L. Community Structure in Industrial SAT Instances. J Artif Intell Res. 2019;66:443-72. DOI: [10.1613/jair.1.11741](https://doi.org/10.1613/jair.1.11741).

A Tableau Calculus for MaxSAT Based on Resolution

Shoulin LI,^a Jordi COLL,^a Djamel HABET,^a Chu-Min LI^{a,b} and Felip MANYÀ^c

^a*Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France*

^b*MIS, Université de Picardie, Amiens, France*

^c*Artificial Intelligence Research Institute (IIIA), CSIC, Bellaterra, Spain*

Abstract. We define a new MaxSAT tableau calculus based on resolution. Given a multiset of propositional clauses ϕ , we prove that the calculus is sound in the sense that if the minimum number of contradictions derived among the branches of a completed tableau for ϕ is m , then the minimum number of unsatisfied clauses in ϕ is m . We also prove that it is complete in the sense that if the minimum number of unsatisfied clauses in ϕ is m , then the minimum number of contradictions among the branches of any completed tableau for ϕ is m . Moreover, we describe how to extend the proposed calculus to solve Weighted Partial MaxSAT.

Keywords. Maximum satisfiability, semantic tableaux, completeness.

1. Introduction

The Satisfiability problem (SAT) is the problem of deciding if there exists a truth assignment for a given propositional formula in conjunctive normal form (CNF) that evaluates the formula to true. An important optimization variant of SAT is Maximum Satisfiability (MaxSAT), which is the problem of finding a truth assignment that minimizes the number of unsatisfied clauses in a multiset of clauses [23]. Note that minimizing the number of unsatisfied clauses is equivalent to maximizing the number of satisfied clauses.

The inference rules applied in SAT are sound if they preserve satisfiability. Nevertheless, such rules are not applicable in MaxSAT because they are usually unsound. Sound MaxSAT inference rules must preserve the minimum number of unsatisfied clauses between the premises and the conclusions. As a consequence, new complete inference systems for MaxSAT have had to be defined [20]. They are MaxSAT extensions of either the resolution rule [33] or semantic tableaux [11,15,34].

This paper presents a new tableau calculus for MaxSAT based on resolution, proves its completeness and defines its extension to Weighted Partial MaxSAT, the case in which some clauses can be declared as hard and soft clauses have an associated weight. The advantage of our calculus is that it can produce shorter proofs than other related approaches in some cases.

Although this work is mainly theoretical, it is worth mentioning that MaxSAT offers a competitive generic problem solving formalism for combinatorial optimization. For example, MaxSAT has been applied to solve optimization problems in real-world domains as diverse as combinatorial testing [1], community detection in complex networks [17],

diagnosis [12], group testing [10], planning [35], routing [31], scheduling [5] and team formation [32], among others. Furthermore, there exist efficient branch-and-bound [21] and SAT-based MaxSAT solvers [4].

The paper is structured as follows. Section 2 defines basic concepts. Section 3 reviews the related work. Section 4 defines a novel MaxSAT tableau calculus based on resolution and proves its completeness. Section 5 defines an extension of the proposed calculus to Weighted Partial MaxSAT. Section 6 gives the conclusions.

2. Preliminaries

A literal is a propositional variable or a negated propositional variable. A clause is a disjunction of literals. A weighted clause is a pair (c, w) , where c is a disjunction of literals and w , its weight, is a natural number or infinity. A clause is hard if its weight is infinity; otherwise, it is soft. The infinity weight is denoted by \top . A weighted partial MaxSAT instance is a multiset of weighted clauses $\phi = \{(h_1, \top), \dots, (h_k, \top), (c_1, w_1), \dots, (c_m, w_m)\}$, where the first k clauses are hard and the last m clauses are soft. A soft clause (c, w) is equivalent to having w copies of the clause $(c, 1)$, and $\{(c, w_1), (c, w_2)\}$ is equivalent to $(c, w_1 + w_2)$. For simplicity, in what follows, we omit weights when all the soft clauses have the same weight.

A truth assignment assigns to each propositional variable either 0 (false) or 1 (true). Weighted Partial MaxSAT, or WPMAXSAT, for an instance ϕ is the problem of finding an assignment that satisfies all the hard clauses and minimizes the sum of the weights of the unsatisfied soft clauses; such an assignment is said to be an optimal assignment.

The Weighted MaxSAT problem, or WMaxSAT, is WPMAXSAT when there are no hard clauses. The Partial MaxSAT problem, or PMaxSAT, is WPMAXSAT when all the soft clauses have the same weight. The (Unweighted) MaxSAT problem is PMaxSAT when there are no hard clauses. The SAT problem, or SAT, is PMaxSAT when there are no soft clauses.

Minimum Satisfiability (MinSAT) is the dual problem of MaxSAT and its goal is to find an assignment that maximizes the number of unsatisfied clauses. The most general extension of MinSAT is Weighted Partial MinSAT, or WPMINSAT, whose goal is to find an assignment that satisfies all the hard clauses and maximizes the sum of the weights of the unsatisfied soft clauses.

3. Related Work

The fact that unit propagation could not be used to simplify CNF formulas in branch-and-bound MaXSAT solvers led to the definition of incomplete resolution-based inference rules for MaxSAT [18,24,25] and of a complete MaxSAT resolution rule [7,8,16]. More recently, the proof complexity community has drawn the attention to MaxSAT resolution with the aim of defining a stronger proof system than SAT resolution. For example, MaxSAT resolution with the split rule (replace clause C with $\neg x \vee C$ and $x \vee C$) produces polynomial-size proofs of the pigeon hole principle, and this does not happen if MaxSAT resolution is replaced with SAT resolution [6,19]. MaxSAT resolution has also been used in a MinSAT branch-and-bound solver [30] and a variable elimination algorithm for MinSAT has been defined in [22,29].

The first tableau calculus for MaxSAT was defined in [27] and then it was extended to non-clausal MaxSAT [13,28]. These works inspired the creation of a complete natural deduction calculus for MaxSAT [9] and tableau calculi for MinSAT [2,3,14,26].

Compared with existing tableau calculi, the calculus of this paper does not need to expand all the clauses in a branch to detect all the possible contradictions. It only needs to expand those clauses containing a complementary literal in another clause of the branch.

4. A MaxSAT tableau calculus based on resolution

We define a MaxSAT tableau calculus and prove its soundness and completeness.

Definition 4.1. *A tableau is a tree with a finite number of branches whose nodes are labelled by either a clause or a box (\square). A box in a tableau denotes a contradiction. A branch is a maximal path in a tree, and we assume that branches have a finite number of nodes.*

Definition 4.2. *Let $\phi = \{\phi_1, \dots, \phi_m\}$ be a multiset of clauses, l a literal, and D and D' disjunction of literals. A tableau for ϕ is constructed by a sequence of applications of the following rules:*

Initialize *A tree with a single branch with m nodes such that each node is labelled with a clause of ϕ is a tableau for ϕ . Such a tableau is called the initial tableau and its clauses are declared to be active.*

Given a tableau T for ϕ and a branch b of T ,

Res-rule *If b contains two active clauses with complementary literals, $l \vee D$ and $\neg l \vee D'$, the tableau obtained by appending a new left branch with two nodes below b labelled with $\neg l$ and D and a new right branch with two nodes below b labelled with l and D' is a tableau for ϕ . Clauses $l \vee D$ and $\neg l \vee D'$ become inactive in b and the added clauses are declared to be active.*

Unit-rule *If b contains an active unit clause l and an active non-unit clause $\neg l \vee D$, the tableau obtained by appending a new left node below b labelled with \square and a new right node with two nodes below b labelled with l and D is a tableau for ϕ . Clauses l and $\neg l \vee D$ become inactive in b and the added non-empty clauses are declared to be active.*

\square -rule *If b contains two active and complementary unit clauses, l and $\neg l$, the tableau obtained by appending a node below b labelled with \square is a tableau for ϕ . Clauses l and $\neg l$ become inactive in b .*

The expansion rules of the previous definition are summarized in Figure 1. Note that all the rules preserve the number of premises falsified by an assignment I in at least one branch and do not decrease that number in the other branch (if any).

Definition 4.3. *Let T be a tableau for a multiset of propositional clauses ϕ . A branch b of T is saturated when no further expansion rules can be applied on b , and T is completed when all its branches are saturated. The cost of a saturated branch is the number of boxes on the branch. The cost of a completed tableau is the minimum cost among all its branches.*

$l \vee D$	l	l
$\neg l \vee D'$	$\neg l \vee D$	$\neg l$
$\neg l$	\square	\square
D	D'	D
<i>Res-rule</i>	<i>Unit-rule</i>	

Figure 1. Tableau expansion rules for non-clausal MaxSAT

Notice that a branch becomes saturated when it does not contain active clauses with complementary literals. We show below that the minimum number of clauses that can be unsatisfied in a multiset of propositional clauses ϕ is m iff the cost of a completed tableau for ϕ is m . Thus, the systematic construction of a completed tableau for ϕ provides an exact method for MaxSAT.

Example 4.4. Figure 2 shows how we can create a tableau, with the previous calculus, to prove that the minimum number of unsatisfied clauses in the multiset of clauses $\{x_1, x_2, \neg x_1 \vee x_3, \neg x_1 \vee \neg x_2 \vee \neg x_3\}$ is one. We first create the initial tableau (the leftmost tableau) and then apply the *Res*-rule to the clauses $\neg x_1 \vee x_3$ and $\neg x_1 \vee \neg x_2 \vee \neg x_3$ (resolving variable x_3), getting as a result the second tableau in the figure. We apply the \square -rule to x_1 and $\neg x_1$ on the leftmost branch and obtain the third tableau. That branch is now saturated because its current active clauses (x_2 and $\neg x_3$) do not contain complementary literals. Then, we apply the *Unit*-rule to the clauses x_1 and $\neg x_1 \vee \neg x_2$ (resolving variable x_1) on the rightmost branch, getting as a result the fourth tableau whose middle branch is saturated (current active clauses: x_2 and x_3). Finally, we apply the \square -rule to x_2 and $\neg x_2$ on the rightmost branch and this branch becomes also saturated (current active clauses: x_1 and x_3). Since the minimum number of boxes among the branches of the last tableau is one, the minimum number of clauses that can be unsatisfied in ϕ is also one.

The advantage of the defined calculus with respect to other MaxSAT tableau calculi [13,27,28] is that it does not need to expand all the active clauses to saturate a branch. It only needs to expand those active clauses containing a complementary literal in another active clause of the branch. This implies, in some cases, that the resulting tableaux have fewer nodes. For instance, the other calculi need to double the number of branches to solve the multiset of clauses of Example 4.4.

4.1. Soundness and completeness

We prove the soundness and completeness of the proposed tableau calculus for MaxSAT. Before presenting the completeness theorem, we prove termination and the soundness of the expansion rules.

Proposition 4.5. *A tableau for a multiset of propositional clauses ϕ is completed in a finite number of steps.*

Proof. We first create an initial tableau and then apply expansion rules in the newly created branches until they become saturated. The *Res*- and *Unit*-rule reduce the number

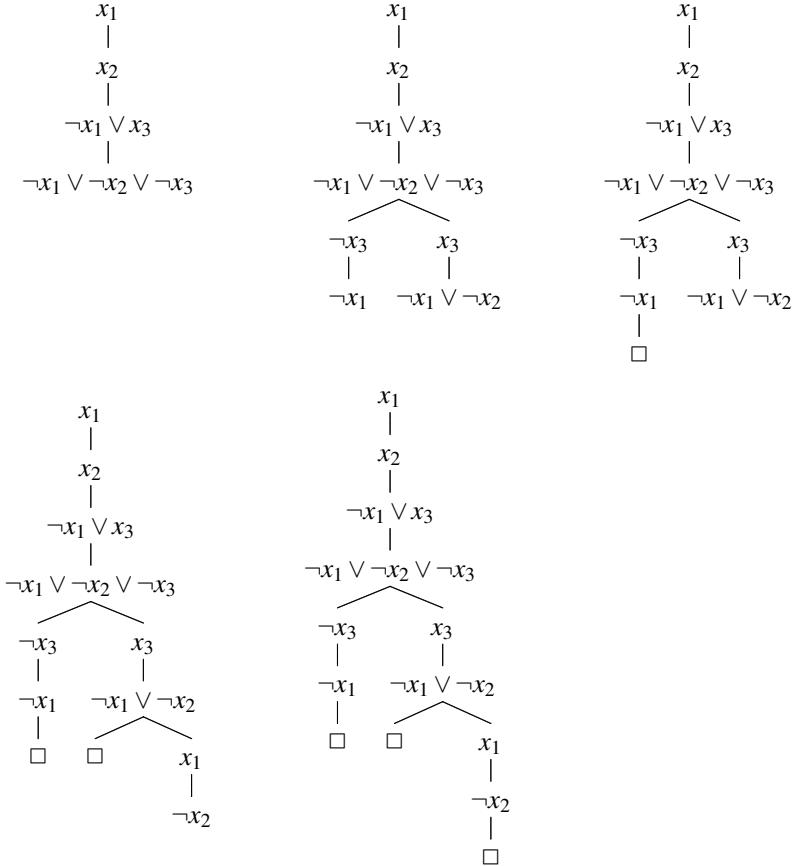


Figure 2. A tableaux for the non-clausal MaxSAT instance $\{x_1, x_2, \neg x_1 \vee x_3, \neg x_1 \vee \neg x_2 \vee \neg x_3\}$.

of connectives. Since we began with a finite number of connectives, these rules can only be applied a finite number of times. The \square -rule inactivates two literals and adds a box. Since we began with a finite number of literals and boxes cannot be premises of any expansion rule, this rule can only be applied a finite number of times. Hence, the construction of any completed tableau terminates in a finite number of steps. \square

Proposition 4.6. *An assignment I falsifies k premises of a Res-, Unit-, and \square -rule iff assignment I falsifies k clauses in one branch of the conclusions of the rule and at least k clauses in the other branch (if any).*

Proof. We prove the result for each rule:

- **Res-rule:** An assignment I satisfies both $l \vee D$ and $\neg l \vee D'$ iff I satisfies either l and D' or $\neg l$ and D . In this case, I satisfies the clauses of one branch and falsifies at least one clause of the other branch. In any other case, I falsifies either $l \vee D$ or $\neg l \vee D'$. If I falsifies $l \vee D$, it falsifies exactly one clause (D) of the left branch and at least one clause of the right branch. If I falsifies $\neg l \vee D'$, it falsifies exactly one

clause (D') of the right branch and at least one clause of the left branch. Hence, the number of unsatisfied clauses is preserved in at least one branch of the rule.

- *Unit-rule:* An assignment I satisfies both l and $\neg l \vee D$ iff I satisfies l and D , and in this case I satisfies the two clauses of the right branch. In any other case, I falsifies either l or $\neg l \vee D$. So, I falsifies the left branch and at least one clause of the right branch. Hence, the number of unsatisfied clauses is preserved in at least one branch of the rule.
- \square -rule: An assignment I either falsifies l or $\neg l$, and satisfies the complementary literal of the unsatisfied literal. Since the single conclusion is a box and denotes a contradiction, I falsifies the same number of clauses in the premises and the conclusion.

□

Theorem 4.7. Soundness & completeness. *The cost of a completed tableau for a multiset of clauses ϕ is m iff the minimum number of unsatisfied clauses in ϕ is m .*

Proof. (*Soundness:*) T was derived by creating a sequence of tableaux T_0, \dots, T_n ($n \geq 0$) such that T_0 is an initial tableau for ϕ , $T_n = T$, and T_i was obtained by a single application of the *Res-*, *Unit-* or \square -rule on an branch of T_{i-1} for $i = 1, \dots, n$. By Proposition 4.5, we know that such a sequence is finite. Since T has cost m , T_n contains one branch b with exactly m boxes and the rest of branches contain at least m boxes. Moreover, the active clauses in every branch of T_n do not contain complementary literals; otherwise, we could yet apply expansion rules and T_n could not be completed. The assignment that sets to true the literals occurring in the active clauses of an optimal branch only falsifies the m boxes and there cannot be any assignment satisfying less than m clauses in a branch of T_n because each branch contains at least m boxes. Therefore, the minimum number of active clauses than can be unsatisfied among the branches of T_n is m .

Proposition 4.6 guarantees that the minimum number of unsatisfied active clauses is preserved in the sequence of tableaux T_0, \dots, T_n . Thus, the minimum number of unsatisfied active clauses in T_0 is also m . Since T_0 is formed by a single branch that only contains the clauses in ϕ and all these clauses are active, the minimum number of clauses that can be unsatisfied in ϕ is m .

(*Completeness:*) Assume that there is a completed tableau T for ϕ that does not have cost m . We distinguish two cases:

(i) T has a branch b of cost k , where $k < m$. Then, T has a branch with k boxes and a satisfiable multiset of active clauses because T is completed. This implies that the minimum number of unsatisfied active clauses among the branches of T is at most k . By Proposition 4.6, this also holds for T_0 , but this is in contradiction with m being the minimum number of unsatisfied clauses in ϕ because $k < m$. Thus, any branch of T has at least cost m .

(ii) T has no branch of cost m . This is in contradiction with m being the minimum number of unsatisfied clauses in ϕ . Since the tableau expansion rules preserve the minimum number of unsatisfied clauses and the branches of any completed tableau only contain active clauses that are boxes or clauses without complementary literals, T must have a saturated branch with m boxes. Thus, T has a branch of cost m .

Hence, each completed tableau T for a multiset of clauses ϕ has cost m if the minimum number of clauses that can be unsatisfied in ϕ is m . □

		$(l \vee D, w_1)$
$(l \vee D, \top)$		$(\neg l \vee D', w_2)$
$\frac{(l, \top) \quad \quad (D, \top)}{\beta\text{-rule}}$		$\frac{(l \vee D, w_1 - w) \quad \quad (\neg l \vee D', w_2 - w)}{(l \vee D, w_1 - w) \quad \quad (\neg l \vee D', w_2 - w)}$
		$\frac{(\neg l, w) \quad \quad (D, w)}{(l, w) \quad \quad (D', w)}$
		where $w = \min(w_1, w_2)$
		<i>Res-rule</i>
(l, \top)		(l, w_1)
$\frac{(\neg l \vee D, w)}{(D, w)}$		$\frac{(\neg l \vee D, w_2)}{(l, w_1 - w) \quad \quad (\neg l \vee D, w_2 - w)}$
		$\frac{(\square, w) \quad \quad (D, w)}{(l, w) \quad \quad (D, w)}$
		where $w = \min(w_1, w_2)$
		<i>Unit-rule</i>
(l, \top)	(l, \top)	(l, w_1)
$\frac{(\neg l, \top) \quad \quad \blacksquare}{(\square, w)}$	$\frac{(\neg l, w) \quad \quad (\square, w)}{(\neg l, w_2) \quad \quad (\square, w)}$	$\frac{}{(\square, w)}$
		$(l, w_1 - w)$
		$(\neg l, w_2 - w)$
		where $w = \min(w_1, w_2)$
		<i>\square-rule</i>

Figure 3. Tableau expansion rules for Weighted Partial MaxSAT

5. A Tableau Calculus for Weighted Partial MaxSAT based on Resolution

Dealing with weighted soft clauses can be understood as collapsing several unweighted MaxSAT inferences into a single inference, because a weighted clause (C, w) can be replaced by w copies of the unweighted clause C . If there are two premises (C_1, w_1) and (C_2, w_2) with different weights ($w_1 \neq w_2$), (C_1, w_1) and (C_2, w_2) become inactive but $(C_1, w_1 - w)$ and $(C_2, w_2 - w)$, where $w = \min(w_1, w_2)$, are added as active clauses (clauses with weight 0 are not added). Then, the conclusions of the inference have weight w . For example, from $(x_1, 1)$ and $(\neg x_1, 3)$ we derive $(\square, 1)$ and $(\neg x_1, 2)$.

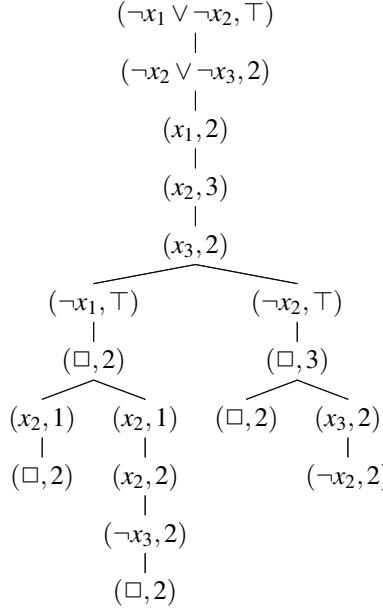


Figure 4. A tableau for the WPMAXSAT instance $\{(\neg x_1 \vee \neg x_2, \top), (\neg x_2 \vee \neg x_3, 2), (x_1, 2), (x_2, 3), (x_3, 2)\}$

When dealing with hard clauses, the inference applied in SAT remains valid in MaxSAT when the premises are hard. Moreover, the detection of a contradiction between two hard clauses implies that we have identified an infeasible solution. In this case, the contradiction is represented by \blacksquare and the branch containing that contradiction can be pruned.

Figure 3 displays the expansion rules of a complete tableau calculus for Weighted Partial MaxSAT. It is formed by the extensions of the *Res-*, *Unit-*, and \square -rules when their premises contain unit hard clauses or weighted soft clauses. In the case in which we have a non-unit hard clause, we can use the β -rule. In our calculus, unit hard clauses are always active while non-unit hard premises become inactive after applying an inference rule. Notice that other inference rules could be used to deal with hard premises but we used the β -rule because it produces a simple and complete calculus.

Example 5.1. Figure 4 displays a tableau for the WPMAXSAT instance $\{(\neg x_1 \vee \neg x_2, \top), (\neg x_2 \vee \neg x_3, 2), (x_1, 2), (x_2, 3), (x_3, 2)\}$. Firstly, we apply the β -rule to $(\neg x_1 \vee \neg x_2, \top)$. Secondly, we apply the \square -rule to $(\neg x_1, \top)$ and $(x_1, 2)$ in the left branch. Thirdly, we apply the \square -rule to $(\neg x_2, \top)$ and $(x_2, 3)$ in the right branch. Fourthly, we apply the *Unit*-rule to $(\neg x_2 \vee \neg x_3, 2)$ and $(x_2, 3)$ in the left branch. Fifthly, we apply the \square -rule to $(\neg x_3, 2)$ and $(x_3, 2)$ in the second leftmost branch. Sixthly, we apply the *Unit*-rule to $(\neg x_2 \vee \neg x_3, 2)$ and $(x_3, 2)$ in the right branch. Since the minimum cost among all the branches is 3, the minimum sum of weights of the unsatisfied soft clauses while satisfying the hard clauses is 3.

Taking into account that a soft clause (c, w) is equivalent to having w copies of clause $(c, 1)$, and $\{(c, w_1), (c, w_2)\}$ is equivalent to $(c, w_1 + w_2)$, we can prove that the previous calculus is complete for WPMaxSAT.

6. Conclusions

We presented a new tableau calculus for MaxSAT based on resolution, proved its completeness and defined its extension to WPMaxSAT. The proposed calculus has the advantage of producing shorter proofs in some cases. Moreover, this work is a step forward to better understanding the logic of MaxSAT. In future work, we plan to extend the calculus to non-clausal MaxSAT, MinSAT and first-order logic.

Acknowledgements: This work has been supported by the French Agence Nationale de la Recherche, reference ANR-19-CHIA-0013-01, and grant PID2019-111544GB-C21 funded by MCIN/AEI/10.13039/501100011033. The last author was supported by mobility grant PRX21/00488 of the Spanish *Ministerio de Universidades*.

References

- [1] C. Ansótegui, F. Manyà, J. Ojeda, J. M. Salvia, and E. Torres. Incomplete MaxSAT approaches for combinatorial testing. *Journal of Heuristics*, (in press), 2022.
- [2] J. Argelich, C. M. Li, F. Manyà, and J. R. Soler. Clause branching in MaxSAT and MinSAT. In *Proceedings of the 21st International Conference of the Catalan Association for Artificial Intelligence, Roses, Spain*, volume 308 of *Frontiers in Artificial Intelligence and Applications*, pages 17–26. IOS Press, 2018.
- [3] J. Argelich, C. M. Li, F. Manyà, and J. R. Soler. Clause tableaux for maximum and minimum satisfiability. *Logic Journal of the IGPL*, 29(1):7–27, 2021.
- [4] F. Bacchus, M. Järvisalo, and M. Ruben. Maximum satisfiability. In A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors, *Handbook of Satisfiability*, pages 929–991. IOS Press, 2021.
- [5] M. Bofill, J. Coll, M. Garcia, J. Giráldez-Cru, G. Pesant, J. Suy, and M. Villaret. Constraint solving approaches to the business-to-business meeting scheduling problem. *Journal of Artificial Intelligence Research*, 74:263–301, 2022.
- [6] M. L. Bonet and J. Levy. Equivalence between systems stronger than resolution. In *Proceedings of the 23rd International Conference on Theory and Applications of Satisfiability Testing, SAT-2020, Alghero, Italy*, pages 166–181. Springer LNCS 12178, 2020.
- [7] M. L. Bonet, J. Levy, and F. Manyà. A complete calculus for Max-SAT. In *Proceedings of the 9th International Conference on Theory and Applications of Satisfiability Testing, SAT-2006, Seattle, USA*, pages 240–251. Springer LNCS 4121, 2006.
- [8] M. L. Bonet, J. Levy, and F. Manyà. Resolution for Max-SAT. *Artificial Intelligence*, 171:240–251, 2007.
- [9] J. Casas-Roma, A. Huertas, and F. Manyà. Solving MaxSAT with natural deduction. In *Proceedings of the 20th International Conference of the Catalan Association for Artificial Intelligence, Deltrebre, Spain*, volume 300 of *Frontiers in Artificial Intelligence and Applications*, pages 186–195. IOS Press, 2017.
- [10] L. Ciampiconi, B. Ghosh, J. Scarlett, and K. S. Meel. A MaxSAT-based framework for group testing. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 10144–10152, 2020.
- [11] M. D’Agostino. Tableaux methods for classical propositional logic. In M. D’Agostino, D. Gabbay, R. Hähnle, and J. Posegga, editors, *Handbook of Tableau Methods*, pages 45–123. Kluwer, 1999.
- [12] D. D’Almeida and É. Grégoire. Model-based diagnosis with default information implemented through MAX-SAT technology. In *Proceedings of the IEEE 13th International Conference on Information Reuse & Integration, IRI, Las Vegas, NV, USA*, pages 33–36, 2012.

- [13] G. Fiorino. New tableau characterizations for non-clausal MaxSAT problem. *Logic Journal of the IGPL*, 2021. doi:10.1093/jigpal/jzab012.
- [14] G. Fiorino. A non-clausal tableau calculus for MinSAT. *Information Processing Letters*, 173:106167, 2022.
- [15] R. Hähnle. Tableaux and related methods. In J. A. Robinson and A. Voronkov, editors, *Handbook of Automated Reasoning*, pages 100–178. Elsevier and MIT Press, 2001.
- [16] F. Heras and J. Larrosa. New inference rules for efficient Max-SAT solving. In *Proceedings of the National Conference on Artificial Intelligence, AAAI-2006, Boston/MA, USA*, pages 68–73, 2006.
- [17] S. Jabbour, N. Mhadhbi, B. Raddaoui, and L. Sais. A SAT-based framework for overlapping community detection in networks. In *Proceedings of the 21st Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Part II, PAKDD, Jeju, South Korea*, pages 786–798, 2017.
- [18] J. Larrosa and F. Heras. Resolution in Max-SAT and its relation to local consistency in weighted CSPs. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-2005, Edinburgh, Scotland*, pages 193–198. Morgan Kaufmann, 2005.
- [19] J. Larrosa and E. Rollon. Towards a better understanding of (partial weighted) MaxSAT proof systems. In *Proceedings of the 23rd International Conference on Theory and Applications of Satisfiability Testing, SAT-2020, Alghero, Italy*, pages 218–232. Springer LNCS 12178, 2020.
- [20] C. Li and F. Manyà. Inference in MaxSAT and MinSAT. In *The Logic of Software*, volume 13360 of *LNCS*, pages 350–369. Springer, 2022.
- [21] C. Li, Z. Xu, J. Coll, F. Manyà, D. Habet, and K. He. Combining clause learning and branch and bound for MaxSAT. In *Proceedings of the 27th International Conference on Principles and Practice of Constraint Programming, CP, Montpellier, France*, volume 210 of *LIPICS*, pages 38:1–38:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.
- [22] C. M. Li and F. Manyà. An exact inference scheme for MinSAT. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina*, pages 1959–1965, 2015.
- [23] C. M. Li and F. Manyà. MaxSAT, hard and soft constraints. In A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors, *Handbook of Satisfiability*, pages 903–927. IOS Press, 2021.
- [24] C. M. Li, F. Manyà, N. O. Mohamedou, and J. Planes. Resolution-based lower bounds in MaxSAT. *Constraints*, 15(4):456–484, 2010.
- [25] C. M. Li, F. Manyà, and J. Planes. New inference rules for Max-SAT. *Journal of Artificial Intelligence Research*, 30:321–359, 2007.
- [26] C. M. Li, F. Manyà, and J. R. Soler. A clause tableau calculus for MinSAT. In *Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, CCIA, Barcelona, Spain*, volume 288 of *Frontiers in Artificial Intelligence and Applications*, pages 88–97. IOS Press, 2016.
- [27] C. M. Li, F. Manyà, and J. R. Soler. A clause tableaux calculus for MaxSAT. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI, New York, USA*, pages 766–772, 2016.
- [28] C. M. Li, F. Manyà, and J. R. Soler. A tableau calculus for non-clausal maximum satisfiability. In *Proceedings of the 28th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, TABLEAUX, London, UK*, pages 58–73. Springer LNCS 11714, 2019.
- [29] C. M. Li, F. Xiao, and F. Manyà. A resolution calculus for MinSAT. *Logic Journal of the IGPL*, 29(1):28–44, 2021.
- [30] C. M. Li, Z. Zhu, F. Manyà, and L. Simon. Optimizing with minimum satisfiability. *Artificial Intelligence*, 190:32–44, 2012.
- [31] Y. Li, S. Lin, S. Nishizawa, and H. Onodera. Mcell: Multi-row cell layout synthesis with resource constrained MAX-SAT based detailed routing. In *IEEE/ACM International Conference On Computer Aided Design*, pages 157:1–157:8, 2020.
- [32] F. Manyà, S. Negrete, C. Roig, and J. R. Soler. Solving the team composition problem in a classroom. *Fundamenta Informaticae*, 174(1):83–101, 2020.
- [33] J. A. Robinson. A machine oriented logic based on the resolution principle. *Journal of the ACM*, 12(1):23–41, 1965.
- [34] R. Smullyan. *First-Order Logic*. Dover Publications, New York, second corrected edition, 1995. First published 1968 by Springer-Verlag.
- [35] L. Zhang and F. Bacchus. MAXSAT heuristics for cost optimal planning. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada*, pages 1846–1852, 2012.

Importance-Performance Analysis in Project Portfolio Management Using an IOWA Operator

Pietro FRONTE^{a,b,1}, Núria AGELL^a Marc TORRENS^a and Daniel BRUGAROLAS^b

^aESADE Business School, Universitat Ramon Llull, Barcelona

^bDigital Office, SEAT S.A.

Abstract. Prioritization of activities is one of the many facets in digital products management. In this study, we propose an alternative point of view on Portofolio Projects Management, in order better understand their relationship between *performance* and *importance* and how resources can be allocated taking into account these two categories. An application of the presented methodology to a real use case in the Digital Department of SEAT S.A. is discussed to show current results and future developments.

Keywords. Information aggregation, Importance-Performance Analysis, IOWA operator, Aggregation function

1. Introduction

Digital businesses can face higher costs of development and R&D but they can also benefit of higher profit margin if the solution become successful compared to traditional businesses [6]. To foster this success, a great contribution is given by the AGILE framework, that allows new digital companies to build an effective and efficient working model since the very beginning.

On the other hand, traditional enterprises are pushed to invest conspicuous amount of resources (workforce, money, time) to keep up with the new digital organizations and convert their portfolio of products including more appealing, convenient and profitable solutions. Automotive manufacturers are examples of traditional enterprises that, until few years ago, were able to generate profit thanks to the bare sale of cars and spare parts. Nowadays, thanks to the digital transformation, they have the opportunity to widen their business channels and increase their profit margins [2].

As for digital startups, AGILE framework helps traditional firms in transforming themselves internally to ride the wave of the digital transformation. In this context, the version of AGILE for large enterprises is usually adopted, namely SAFe (Scaled Agile Framework for Enterprise). SAFe encompasses a set of practices, roles, duties that helps enterprises in managing and developing digital products portfolio [4].

¹Corresponding Author: Pietro Fronte, ESADE Business School, pietro.fronte@esade.edu

Prioritization of activities is one of the critical tasks in digital products management that is regulated and supported by SAFe with different, easy to implement, techniques. The most used one in prioritizing is the Weighted Shortest Job First (WSJF) method [7]. The WSJF is a well suited method to prioritize sub-activities (Features, User Stories) of wider projects/program. It takes into account the business value each improvement will provide once developed.

This WSJF methodology has been adopted also by the Digital Office of SEAT S.A. when prioritizing the set of portfolio projects that the company is going to pursue in the mid term (2022-2026). The objective of the prioritization task was to set priorities of the department at portfolio level, and obtain a first hint on which projects could have been discarded. For the sake of the purpose, the WSJF had been slightly adapted to meet the criteria available relatively to the original set of project. The result, however, did not meet the expectation of the management that negotiated some changes in the final ranking to reach a wider consensus.

The goal of this study is therefore to propose an alternative point of view on Portfolio Projects Management, in order better understand their relationship between *performance* and *importance* and how resources can be allocated taking into account these two categories.

The rest of the paper is organized as follows. In Section 2, the theoretical preliminaries are provided. Section 3 presents a real case application of the framework within the digital department of an automotive company. Finally, some conclusions and future research are discussed.

2. Preliminaries: Theoretical approach

In this section, we briefly introduce the necessary concepts and the methodology for the approach considered. This approach adopts a hybrid solution, merging Multi-Criteria Decision Aiding, Importance Performance Analysis and Information Aggregation.

We consider two different categories of variables: variables related to the economic performance and variables related to the strategic importance of the activities.

The original setup therefore includes:

- The set of alternatives (activities or projects) to rank $A = \{a_1, a_2, \dots, a_n\}$
- The set of economic variables $V_e = \{v_{e1}, v_{e2}, \dots, v_{em}\}$
- The set of strategic variables $V_s = \{v_{s1}, v_{s2}, \dots, v_{sz}\}$

For each set of variables, Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is adopted to assign to each alternative a global value of economic performance and a global value of strategic importance, $P = \{p_1, p_2, \dots, p_n\}$ and $I = \{i_1, i_2, \dots, i_n\}$ such that $p_j, i_j \in [0, 1]$ for each $j = 1, \dots, n$ [3].

Activity	Economic performance value	Strategic importance value
a_1	p_1	i_1
a_2	p_2	i_2
...
a_n	p_n	i_n

Table 1. TOPSIS similarity values for Economic performance and Strategic importance

For a qualitative and visual interpretation of the values calculated, alternatives are reported into an Importance-Performance Analysis diagram [1], adopting the Economic performance values as x-coordinates and the Strategic importance as y-coordinates.

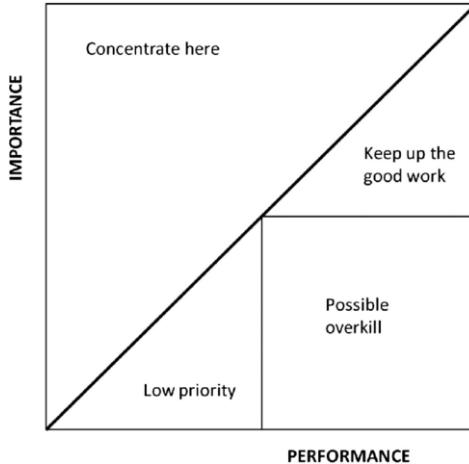


Figure 1. Importance-Performance Analysis diagram partition introduced in [1]

Unlike the traditional Importance-Performance Analysis, where the ordinal ranking position of alternative is reported into the diagram, the Importance/Performance similarity measures from TOPSIS are plotted.

After computing the two different, independent, values of Economic performance and Strategic importance for each alternative, an IOWA-based operator, presented in [5], is adopted to retrieve a global evaluation index for the portfolio: the \mathcal{G} -index. The objective of this index is to provide a global score of importance vs performance status of the whole portfolio, with a focus on under-performing alternatives.

By definition [8], an IOWA operator of dimension n is a function $\Phi : (\mathbb{R} \times \mathbb{R})^n \rightarrow \mathbb{R}$ such that

$$\Phi((u_1, x_1), \dots, (u_n, x_n)) = \sum_{i=1}^n w_i x_{\sigma(i)} \quad (1)$$

where

- $\{u_1, \dots, u_n\}$ is the order-inducing variable of the IOWA operator
- $\{x_1, \dots, x_n\}$ is the argument variable of the IOWA operator
- $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ is a permutation such that $u_{\sigma(i)} \geq u_{\sigma(i+1)}$ for $i = 1, \dots, n-1$
- $\{w_1, \dots, w_n\}$ is the set of weights such that $w_i \in [0, 1]$ and $\sum_{i=1}^n w_i = 1$.

In this specific case, and as employed in [5], n is the number of alternatives considered whilst u_i and x_i are replaced by p_j and i_j , respectively.

The Importance vector of the alternatives assumes the role of Order-inducing variable while the Performance vector takes the role of argument variable. Consequently, the Importance-Performance (IP) vector of the initial set of alternatives becomes:

$$\left((i_1, p_1), \dots, (i_n, p_n) \right) \quad (2)$$

with the tuple (i_1, p_1) belonging to the most important alternative with its associated performance and (i_n, p_n) related to the least important one with its corresponding performance value.

As previously stated, the \mathcal{G} -index focuses primarily on the under-performing alternatives, defined as items whose Performance value p_j is smaller than the Importance counterpart i_j . To account for this aspect, the Non-Negative Performance-Importance vector is calculated as

$$DV = (X_1, \dots, X_n) \text{ with } X_j = \max(p_j - i_j; 0) \text{ for } j = 1, \dots, n \quad (3)$$

The \mathcal{G} -index is finally defined as:

$$\mathcal{G}(X_1, \dots, X_n) = \sum_{j=1}^n w_j X_j \quad (4)$$

where the weights are represented by the normalized Importance vector, i.e. $w_j = \frac{i_j}{\sum_{j=1}^n i_j}$ for all $j = 1, \dots, n$, so $w_j \in [0, 1]$ for all $j = 1, \dots, n$ and $\sum_{j=1}^n w_j = 1$. In addition, $w_j X_j$ represent the marginal contribution of the j^{th} alternative to the \mathcal{G} -index, with the contribution's magnitude varying according to its importance value i_j .

3. Application to a real case: SEAT S.A.

The proposed methodology and framework has been applied to a real case problem within the Digital Office of SEAT S.A. The problem setup involves a set of 14 activities to be ranked taking into account 7 variables, categorized into economic variables (3) and strategic variables (4). The aforementioned 14 activities will be pursued by the Digital Office in the mid-term (2022-2026). For this reason, the values available for the economic variables are estimations provided by the management, while the strategic variables reflect the current importance that a project have for the department. Projects' names, variables' values and names have been omitted due to confidential information.

	Economic performance variables			Strategic importance variables			
	v_{e1}	v_{e2}	v_{e3}	v_{s1}	v_{s2}	v_{s3}	v_{s4}
$Project_1$	$v_{e1,1}$	$v_{e2,1}$	$v_{e3,1}$	$v_{s1,1}$	$v_{s2,1}$	$v_{s3,1}$	$v_{s4,1}$
...
$Project_i$	$v_{e1,i}$	$v_{e2,i}$	$v_{e3,i}$	$v_{s1,i}$	$v_{s2,i}$	$v_{s3,i}$	$v_{s4,i}$
...
$Project_{14}$	$v_{e1,14}$	$v_{e2,14}$	$v_{e3,14}$	$v_{s1,14}$	$v_{s2,14}$	$v_{s3,14}$	$v_{s4,14}$

Table 2. Original dataset composition.

Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is applied to first identify, for the two categories of variables (economic and strategic), the Ideal and Anti-Ideal solution and subsequently calculate the relative Euclidean distance of each alternative with respect to the two endpoints.

The final values are:

	Economic Performance	Strategic Importance
Project1	0.160	0.447
Project2	0.615	0.588
Project3	0.462	0.200
Project4	0.486	0.700
Project5	0.100	0.313
Project6	0.000	0.447
Project7	0.816	0.413
Project8	0.256	0.700
Project9	0.454	0.800
Project10	0.226	0.700
Project11	0.000	1.000
Project12	0.130	0.800
Project13	0.081	0.656
Project14	0.044	0.568

Table 3. Similarity values for Economic Performance and Strategic Importance Obtained via TOPSIS

Similarity values are then reported into an IPA diagram to visually reflect the current importance-performance trade-off of the portfolio. The current diagram visually shows a condition of general under-performing portfolio. The majority of portfolio items fall within the area "Concentrate here" of the considered partition of the IPA diagram while only a single item sits within the "Keep up with the good work" section.

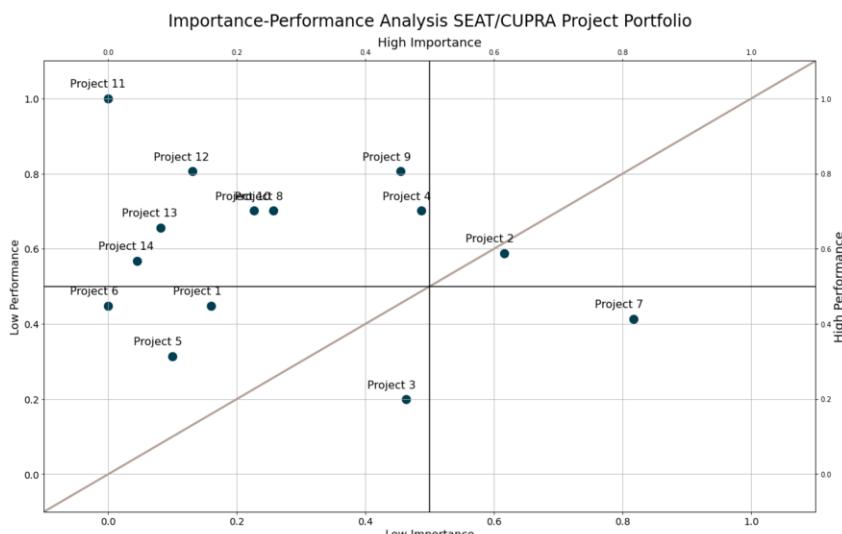


Figure 2. TOPSIS similarity values reported in IPA diagram

As previously defined in Equations 3 and 4, the Non-Negative Difference values X_j , the weights w_j and the marginal contributions $X_j * w_j$ for $j = 1, \dots, n$ are calculated and reported in Table 4.

	Economic Performance	Strategic Importance	Non-Negative DV	Weight	Marginal Contribution \mathcal{G}
Project1	0.160	0.447	0.287	0.054	0.015
Project2	0.615	0.588	0.000	0.071	0.000
Project3	0.462	0.200	0.000	0.024	0.000
Project4	0.486	0.700	0.214	0.084	0.018
Project5	0.100	0.313	0.213	0.038	0.008
Project6	0.000	0.447	0.447	0.054	0.024
Project7	0.816	0.413	0.000	0.050	0.000
Project8	0.256	0.700	0.444	0.084	0.037
Project9	0.454	0.800	0.346	0.096	0.033
Project10	0.226	0.700	0.474	0.084	0.040
Project11	0.000	1.000	1.000	0.120	0.120
Project12	0.130	0.800	0.670	0.096	0.064
Project13	0.081	0.656	0.575	0.079	0.045
Project14	0.044	0.568	0.524	0.068	0.036

Table 4. Non-negative difference values, weights, and marginal contributions for each alternative.

Following the results shown in Table 4, the original IPA diagram shown in Figure 2 is modified taking into account the weight that each project has with respect to the entire portfolio set 3.

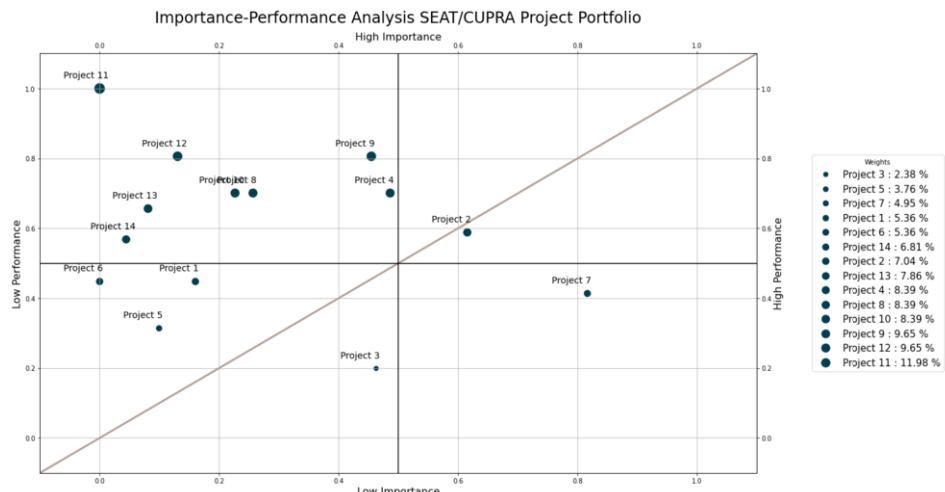


Figure 3. TOPSIS similarity values reported in IPA diagram with scatter points' size linked to alternatives' weights

Finally, to quantify the degree of under-performance, the \mathcal{G} -index as introduced in [5] is applied. According to the values of Table 4, the overall \mathcal{G} -index for this portfolio

is 0,441. Note that accordingly to our approach, the optimum would be achieved when $\mathcal{G}(a_1, \dots, a_n) = 0$.

4. Conclusions and future research

This paper proposes an alternative point of view on Portfolio Projects Management, in order better understand the relationship between performance and importance and how resources can be allocated taking into account these two categories. The final objective is to be able to identify, and provide, the right action(s) to take in terms of portfolio performance improvement, given a certain amount of resources available. The presented research is still an on-going research, for this reason there are still many features of the method that should be carefully revised and aligned with the Decision Makers.

The next challenges to be solved are the optimization problem related to the allocation of resources minimizing the \mathcal{G} -index as objective function and, the improvement of the IOWA operator weights contributing to the final \mathcal{G} -index calculation. Also, we plan to analyze and compare results with other OWA operators.

References

- [1] Abalo, Javier, Jesús Varela, and Vicente Manzano. "Importance values for Importance–Performance Analysis: A formula for spreading out values derived from preference rankings." *Journal of Business Research* 60.2 (2007): 115-121.
- [2] Athanasopoulou, Alexia, et al. "The disruptive impact of digitalization on the automotive ecosystem: a research agenda on business models, platforms and consumer issues." *Bled econference*. 2016.
- [3] Hwang, Ching-Lai, and Kwangsun Yoon. "Methods for multiple attribute decision making." *Multiple attribute decision making*. Springer, Berlin, Heidelberg, 1981. 58-191.
- [4] SAFe 5.0 Framework, <https://www.scaledagileframework.com/>, Scaled Agile Framework, Date accessed May 30, 2022, Date published January 19, 2022
- [5] Sayeras, Josep M., et al. "A measure of perceived performance to assess resource allocation." *Soft computing* 20.8 (2016): 3201-3214.
- [6] Sundaram, Rammohan, Dr Sharma, and Dr Shakya. "Digital transformation of business models: a systematic review of impact on revenue and supply chain." *International Journal of Management* 11.5 (2020).
- [7] WSJF, <https://www.scaledagileframework.com/wsjf/>, Scaled Agile Framework, Date accessed May 30, 2022, Date published April 09, 2021
- [8] Yager, Ronald R., and Dimitar P. Filev. "Induced ordered weighted averaging operators." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29.2 (1999): 141-150.

This page intentionally left blank

Sentiment Analysis and Text Analysis

This page intentionally left blank

Streamlining Text Pre-Processing and Metrics Extraction

Elena Álvarez-García ^a and Daniel García-Costa ^a and Francisco Grimaldo ^{a,1}

^aDepartment of Computer Science, University of Valencia, Spain

Abstract. Natural Language Processing involves reshaping and refining data sets into data that can be used for analysis, ensuring that the data is well formatted. The efficiency gap of data scientists spending most of their time preparing data is an opportunity for the technology sector to work on solutions to the problem. For this reason, a web tool has been developed that is capable of, on the one hand, speeding up the text cleaning process and, on the other hand, facilitating the extraction of metrics by analyzing and processing the texts through customized dictionaries in LIWC format, uploaded by the users themselves, and through sentiment analysis. All this, from a single interface that allows the user to customize the whole pipeline offering different modules for pre-processing and metrics extraction in order to be a solution to facilitate, streamline and automate the whole process.

Keywords. Text pre-processing, NLP, Text Mining, Metrics, Web tool

1. Introduction

Data quality is one of the main factors in data science, and clean data is important for creating great learning models and preparing data for latter analysis. The results of a CrowdFlower survey [2] of 16,000 data scientists reveal that time spent on pre-processing is one of the main obstacles. Compared to other tasks such as data mining and creating training sets, cleaning and organising data consumes around 60% of their time. According to Anaconda's 2020 annual study [1], the efficiency gap of data scientists spending most of their time preparing data for further analysis has been identified as an opportunity for the technology sector to work on solutions to the problem.

In Natural Language Processing (NLP), text processing refers to the practice of cleaning and preparing text data. To apply NLP techniques it is crucial that the corpus is well pre-processed. For example, data pre-processing is an essential step for the construction of machine learning models, whose outcomes can differ depending on it.

Pre-processing involves reshaping and refining texts into data that can be used for analysis, ensuring that the data is well formatted and follows a set of rules so that it can be understood and processed by machines [3]. Texts often contain characters such as punctuation or stop words that do not provide information and increase the complexity of the analysis. Thus, it is recommended to remove as much noise as possible before analyzing the data, in order to obtain clean data easy to process and to analyse later.

¹Corresponding Author: Francisco Grimaldo, Department of Computer Science, University of Valencia, Spain; E-mail: francisco.grimaldo@uv.es.

Another relevant issue while analyzing text is text mining. It is based on different techniques and technologies that explore large amounts of data automatically or semi-automatically. Text mining seeks to discover information that was not present in a specific way, and text pre-processing is also a key step to successfully apply these techniques [6].

Nowadays there are multiple applications focused on text pre-processing (e.g. cleaning) and other focused on text mining (e.g. sentiment analysis), but they usually work independently. As for applications focused on text cleaning, two of the best known on the market are: Trifaca Wrangler² and OpenRefine³, both of them try to facilitate text pre-processing. The number of web tools capable of extracting sentiment from text is very high, being MonkeyLearn⁴ one of the most renowned over the last year. This application hosts a set of text analysis tools, including sentiment analysis and keyword extraction. Unfortunately most of these applications are not free and do not provide a full set of tools that combine different types of techniques.

Unlike the aforementioned applications, the proposed web tool seeks to bring together both NLP techniques and text mining techniques using different Python libraries that allow the pre-processing of texts and the extraction of metrics for datasets in the simplest, fastest and most automatic way possible, having the aim of reducing the time spent in pre-processing and text mining, making data scientist's work easier and faster.

2. Proposal

The proposed web tool tries to make the pre-processing of text faster and to facilitate the extraction of metrics by applying dictionary based techniques, sentiment analysis and others, combining both methods for pre-processing and text mining in one single tool.

Bringing together these two types of techniques in a single tool will make it easier, for researchers not so familiar with specialised libraries, to perform these tasks in languages like Python, by offering an interface that allows them to automate and customise text pre-processing.

In order to obtain a scalable and modular web tool which can increase the number of features over time, it has been developed implementing different micro-services that provide each of the different functionalities.

The operation of the tool is straightforward. Users upload datasets in CSV format, they personalize the pipeline of the process and choose which order and characteristics to apply to each text column. As a result, the dataset is enriched with new columns depending on the customized pipeline. For example, if cleaning is the only technique selected, a new column with clean text is added, whereas if a custom dictionary extractor is activated, the user will get one new column for each category in the selected dictionary.

3. Use Case

By way of example, we present a use case of pipeline that provides the user with 4 different techniques, each one implemented in a different module (micro-service). Two of

²<https://www.trifacta.com/>

³<https://openrefine.org/>

⁴<https://monkeylearn.com/>

them are focused on text-preprocessing techniques (namely the Cleaning Module and the Translation Module) and the other two in text mining (namely the Custom Dictionary Module and the Emotions Module). The web tool then allows the user to configure the list of tasks and customize the characteristics of each module.

Cleaning Module: The aim of this module is to obtain clean texts that can be later parsed and analyzed without errors. Users can select which characteristics apply in this process such us: convert to upper or lower case, remove stop-words, remove punctuation marks, remove special characters, trim and remove multiple spaces or stem.

Translation Module: The purpose of this module is to translate texts into other languages. From the different Python libraries that perform this task, we use Google Trans, the free Python library that implements the Google API. This library works by making calls to the Google Translate API and allows to translate texts into all the languages supported by the Google API.

Custom Dictionary Module: This module is responsible for making quantitative analysis of texts by checking for the presence of words given by a dictionary in the LIWC⁵ (Linguistic Inquiry and Word Count) format uploaded by the user. LIWC is a well-known format available for a large number of existing dictionaries and its ease to use by people with less specialised knowledge. Using LIWC, words can be counted and grouped in meaningful categories [5]. In this way, it is possible to obtain a value for each of the dimensions defined by the dictionary. These metrics are extracted by means of the liwc⁶ library in Python, which parses dictionaries and counts the occurrences of words within the text. We then obtain the percentage with respect to the total number of words in the text. The outcome of this module adds a new column to the dataset for each of the categories specified in the dictionary.

Emotions Module: Sentiment analysis, also known as opinion mining, is the field of study that analyses people's opinions, feelings, evaluations, attitudes and emotions towards entities and their attributes expressed in written text [4]. This module extracts sentiments form texts written in English and Spanish using the libraries Lexmo⁷ and pysentimiento⁸, respectively.

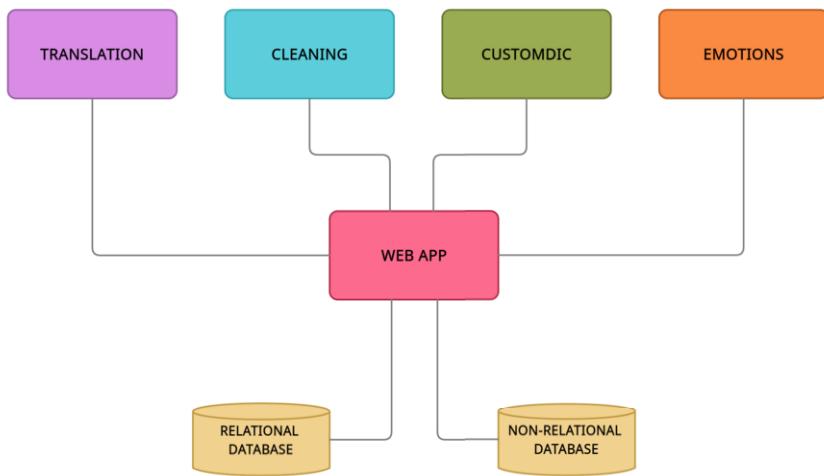
The following figure shows the architecture of the web tool using the 4 modules previously explained. Each of these modules is an independent micro-service that communicates with the application that displays the web interface and from which the user interacts. This web application is in charge of the persistence of all the data, using a relational database for user information and a non-relational one for storing the datasets. The use of a non-relational database allows us to deal with different data structures in the provided datasets.

⁵<https://www.liwc.app/>

⁶<https://github.com/chbrown/liwc-python>

⁷<https://github.com/dinbav/LeXmo>

⁸<https://github.com/pysentimiento/pysentimiento>

Figure 1. Application architecture

4. Conclusions

The tool allows the application of NLP and text mining techniques such as cleaning and extraction of metrics such as sentiment analysis through an easy-to-use interface, obtaining machine and human understandable information. The entire processing pipeline can be customised by the user, helping scientists from other disciplines that do not have much knowledge in these pre-processing, and also for more experienced people to speed up certain tests or parts of their work.

The application is currently in alpha phase and is expected to be made accessible to anyone when it is in a more stable state.

References

- [1] Anaconda. Moving from hype toward maturity 2020 state of data science, 2020.
- [2] CrowdFlower. Datascience report, 2016.
- [3] Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, Freeson Kaniwa, and Kutlwano Sedimo. Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, 6(3):207–210, 2016.
- [4] Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38, 2015.
- [5] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.
- [6] Yu Zhang, Mengdong Chen, and Lianzhong Liu. A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 681–685, 2015.

Drake or Hen? Machine Learning for Gender Identification on Twitter

Arnault GOMBERT^{a,1}, and Jesus CERQUIDES^b

^a *Citibeats, Barcelona, Spain*

^b *Artificial Intelligence Research Institute (IIIA), CSIC, Cerdanyola 08193, Spain*

Abstract. Social media offers an invaluable wealth of data to understand what is taking place in our society. However, the use of social media data to understand phenomena occurring in populations is difficult because the data we obtain is not representative and the tools which we use to analyze this data introduce hidden biases on characteristics such as gender or age. For instance, in France in 2021 women represent 51.6% of the population [1] whereas on Twitter they represent only 33.5% of the french users [2]. With such a difference between social networks user demographics and real population, detecting the gender or the age before going into a deeper analysis becomes a priority. In this paper we provide the results of an ongoing work on a comparative study between three different methods to estimate gender. Based on the results of the comparative study, we evaluate future work avenues.

Keywords. machine learning; deep learning; bias; NLP; gender; Twitter

1. Introduction

Social networks provide a rich amount of real time data that social scientists analyze to find actionable insights. For instance [3] investigates Chilean citizens perception of transport, [4] looks at how social media influenced the 2016 presidential campaign in the United States or [5] focuses on emotions and narratives with respect to gender. The inherent bias carried by social networks illustrates the need to identify gender for social scientists when analyzing the networks before drawing any assumption or conclusion.

Twitter has an easy access to their data using their own API². Besides, the community has focused mainly a lot on the Twitter data. For instance [3,5,6,7,8,9,10,11] based their research only on tweets and did not address other social networks. As argued by Steinert-Threlkeld [12]: “Twitter presents an ideal combination of size, international reach, and data accessibility that make it the preferred platform in academic studies”. In this work we tackle the problem of gender identification on Twitter.

There are two main research lines dealing with gender identification on Twitter. The first one deals with the capability to identify the gender of an author based fundamentally on the texts generated by the user. This research line is very well summarized by

¹Corresponding Author: Arnault Gombert, Citibeats, Barcelona, Spain. E-mail:agombert@citibeats.com

²<https://developer.twitter.com/en/docs/twitter-api>

Ikae and Savoy in [13] and has as corner stone the author profiling task at PAM-CLEF competitions [14,15,16,17,18,19,20].

The second one does also use the metadata related to the tweet and to its author that can be obtained via the Twitter API. These tweet and author metadata are used as inputs of the gender identification model. Quite some work has been devoted to identify gender just based on the name since the seminal work of Michael [21]. Among them, we highlight Demographer [22,23] and [24]. Other works [3] tackle the problem with classical machine learning techniques converting *names*, *description* and/or *username* in a *bag-of-words* representation through word or character *n-grams*. We also see a development of gender identification through deep-learning methods, for instance [8,9] added a computer vision component to process the user profile picture to improve results. And [9] processed *names*, *description* and *username* through a sequential model. More recently we saw models with pre-trained BERT-architecture [25] like in [6] to determine the user's gender. And [10] focused on the difference between using classical machine learning and deep learning methods.

One of the main difficulties is to build one model dealing with several languages at the same time, a lot of works [3,4,5,6,10,22,7] focused on one or two languages only. Wang et al. [9] are the first ones to tackle the problem in a multilingual paradigm, processing 32 different languages. Nevertheless, the possibility to evaluate the models and create a benchmark is complicated due to the lack of labeled data in those several languages.

A more fine grained understanding of the typology of Twitter users is provided when we are able to discern between human accounts from accounts coming from organizations as in [26]. This line of work is continued in [9] and we do also focus our work in finding out if users are real humans or institutional account such as a company or official pages.

In this work we present an initial set of results of our efforts to set up a production multilingual gender identification system on Twitter. Our contributions are

- (i) an efficient and scalable methodology to create multilingual datasets with Twitter users soft labeling, and
- (ii) a baseline evaluation of the quality of three models for gender identification based on this data.

Following this line of work, in the near future, we plan to

- (i) open a labeled dataset that could be used as test set to benchmark any gender machine learning model,
- (ii) open our model via one API like in [11] or in the HuggingFace platform[27], and
- (iii) quantify our model bias regarding gender occupations like in [28] and mitigate it if necessary.

2. Methods

2.1. A methodology for building multilingual Twitter user demographics pseudo labeled datasets

We are interested in designing a methodology that can be used to create *reasonably good* datasets of Twitter users annotated as either *institution*, *man*, or *woman*. The

methodology is required to be easily adaptable to any new language. Labeling data is time consuming and as we have access to a vast amount of unlabeled data on Twitter, we focus on soft-labels to fast annotate a large amount of data. We design a procedure that fits all languages.

The main idea of our procedure is to combine different soft-labelers to create more robust datasets. The first one uses self-reporting info as in [3,9,6]. For instance users that have in their description *father* and *grandpa* or *official account* are soft-labeled respectively as *man* or *institution*. We provide a list of self-reporting words or expressions that can be considered as self-reporting³. The second soft-labeler is inspired from [22], we look, for each language, at a list of first names and its number of occurrences for men and women, then we applied it on Twitter names to get an estimator of the gender. For instance in the US we have access to the distribution of gender by name⁴. We also used a third soft-labeler based on dependency parsing and *part-of-speech tagging* to detect if the user *description* if the user is expressing it/her/himself as a person, an institution or a group of persons, for instance like "*I love pancakes*" would never refer to an institution whereas a description like "*Our firm helps ...*" would. Then we combine those soft labels to identify better the user demographic with a majority vote across all soft-labelers. In case of ties, we do not include the data into the final set.

2.2. A first soft-labeled dataset

At first, we applied the procedure on five different languages: *Catalan*, *English*, *French*, *Portuguese* and *Spanish*. We sampled Twitter from 2016 to 2020 to get 3 million users in total. Table 1 describes the demographic labels in our datasets after applying our soft-labelers and balancing our dataset in order to have a 50% of the users being institutions, 25% being men and 25% being women. A rough estimate of the equivalent amount of data Wang et al. [9] would have used for training with only 5 languages instead of 32 would be 2.27M users. Thus, our dataset is about a 5% in terms of size of Wang et al. dataset. Table 2 describes the distribution of data across the different languages. Both tables also include the numbers for the split of the total dataset into a 70% for training and a 30% for testing stratified by language.

Dataset	Total	organization	male	female
Total	106810	53405	26703	26702
Train	71562	35781	17891	17890
Test	35248	17624	8812	8812

Table 1. Distribution of collected data by demographic labels.

2.3. Experimental Design

Our model considers as inputs only three of the metadata texts associated with a Twitter user: its *name*, its *description* and its *username*. And it considers a two variables outputs (y_1, y_2) with $y_1 = 1$ if the user is an institution, and 0 otherwise. As for y_2 we have that

³https://drive.google.com/drive/u/0/folders/1JP-x0lwzLU8Ue_jRyXGtPRdWrjRWJ3Eu

⁴<https://www.ssa.gov/oact/babynames/limits.html>

Dataset	Catalan	English	French	Portuguese	Spanish
Total	6560	28606	11173	13890	46581
Train	4390	19220	7411	9434	31107
Test	2170	9386	3762	4456	15474

Table 2. Distribution of collected data across languages

$y_2 = 1$ if the user is a woman, and $y_2 = 0$ if the user is a man and otherwise it does not matter.

We created three different models. The objective is to explore the capacities of a baseline model without machine learning, a classical machine learning approach with a *bag-of-words* representation and a deep learning model in the line of the one presented by Wang et al. [9].

The name may be a strong factor for defining the final gender. But as we want our model to also learn from features present in the description we would like to attenuate the importance given to the name. In order to smooth the name occurrence in the training set we decided to mask the name with a probability inversely proportional to its frequency in the training set. Thus the most frequent name such as Thomas or Juliet will be learnt but the model will also focus more on the description. Furthermore in order to avoid overfitting from the gender masks we used in soft-labels, we decided to mask them with a probability of 80% the gendered marks we have identified with our soft-labelers to avoid overfitting. The loss functions are also adapted as in [29] to consider the second output if and only if the observation is a human.

2.3.1. Baseline model

The baseline model is based on an *a priori* gender marked words such as *mistress* or *waiter* to identify gender. To detect if the user is a human being, we look if the name of the user contains a first name from the ones we gathered with the gender distribution in each language for our second soft-labelers. If the name does not contain any of the first name then we considered the user as an institution otherwise a human being. Then to differentiate the gender we used the baseline created in [30]. This baseline uses weights from a regression model to detect gender on social networks. We applied this regression on the concatenated name and description.

2.3.2. Machine learning model

The second model is based on a simple pipeline combining a *term-frequency times inverse document-frequency* (TF-IDF) representation and a logistic regression classifier. The TF-IDF looks at *unigrams* and *bigrams*. It has as parameters a list of predefined stop-words in each language and considers only tokens appearing at least twice in the training set. The logistic regression had a *L1-regularizer* penalty.

2.3.3. Deep learning model

Our model is inspired from [9]. We decided to get rid of the computer vision part and to focus on a text inputs only: *name*, *username* and *description*. First we train for each input an independent bi-directional long short term memory recurrent neural networks (bi-LSTM RNN) model to predict the gender. For instance, only with the input name, we

Model	F1 - Institutions / People	F1 Men/Women
Wang et al. (2019)	89.90	91.8
Baseline	63.5	57.0
BoW + Logit	78.4	87.3
Ours	78.9	91.5

Table 3. Results of different implementations for 5 languages together

train a bi-LSTM RNN model to predict if the name is more likely to belong to a man, a woman or an organization. Second, we get the three trained models back, discard the final softmax layer of each model, concatenate the last layer of each model together and add a new softmax layer on top of it. We train this architecture in two steps. During the first one, the warm-up step, we freeze all layers but the the final softmax layer and train the model. Then we unfreeze all layers and train all layers together.

We chose bi-LSTM RNN [31] to be aligned with [9] and compare our two ways of constructing our training sets. LSTMs advantage is that they learn long-distance relationships between the inputs.

We trained for each of the inputs a bi-LSTM with character and word embeddings. The character embedding representation enables to represent shared linguistic patterns across languages as suggested in [32]. Besides character embeddings representation with LSTM is pretty efficient to detect morphologically rich language as exposed in [33]. Both embeddings, word and character, will then represent better the meaning of the inputs in a multilingual paradigm.

3. Experimental results

First, we look at the institution detection in 3 on the left. Our best model is not so far from [9]. We reach 87.7% of their results so far with 4.7% of the data volume and without using profile pictures. The deep learning model clearly outperforms the baseline but the classical machine learning model reaches close results. We should increase our training set volume to get the full potential of the deep learning architecture to get results as in [10]: a significant margin between deep learning and classical machine learning methods. Nevertheless, our methodology goes in the good direction as we reach similar results with much lower data.

<i>n</i> -grams	P[male]	P[female]
cris	0.25	0.75
ina	0.11	0.89
nat	0.39	0.61
isco	0.98	0.02

Table 4. Empiric Probabilities of some *n*-grams to belong to a man or a woman

When we look at the gender differentiation in 3 on the right, we see that our model is almost at the level of [9]. We reach 99.7% of the results with only 4.7% of the data volume they use and without using profile pictures. We also outperform the two other methodologies: the baseline and the classical machine learning by a significant margin.

Indeed the difference between our method including recurrent neural network and the bag of words representation with a logistic regression is higher than 4 points. It confirms results from [10] about significant improvement when detecting gender for deep learning model compared to classical machine learning models.

We also have evidence that our deep learning model learnt well how to differentiate women and men. First, we have computed the empiric probabilities associated with some features such as diminutive names: *cris* has a probability of 75% to be associated with women, as it can also be used by men. We show some examples in Table 4.

4. Conclusions and future work

In our ongoing work we tackle the problem of differentiating people from institution and identifying the gender of real users simultaneously. We propose a soft-labeling based methodology to build our data sets by combining three different methods to tag the collected data from Twitter. In the future, we will add additional soft-labelers, for instance existing classifiers such as [19] or [22] to leverage existing knowledge. The soft-labelers can be adapted to a lot of languages in a relatively small amount of time. We also plan to experiment with calibrated soft-labeling using the methods presented in [34].

We see that our methodology enables to reach good results with a much lower data volume. Besides for gender detection we are almost at the same level than [9] when working on 5 languages and with only text inputs. We also see that the character embeddings paradigm catches morphological variations in a multilingual framework. We will work on adding more data to leverage the deep learning potential, more languages and we will also add the text of the tweet as input as it can carry gender information and also patterns associated to institutions.

We will also provide an ablation study on the different inputs to quantify the importance of each one in the decision. In parallel we will quantify our model gender bias regarding occupations. Our plan is to also provide different labeled test set that anyone can use to benchmark her/his model and to open-source our final model to improve reusability.

References

- [1] La fécondité se maintient malgré la Pandémie de Covid-19;. Available from: <https://www.insee.fr/fr/statistiques/6024136>.
- [2] Kemp S. Digital in France: All the statistics you need in 2021 - DataReportal – global digital insights. DataReportal – Global Digital Insights; 2021. Available from: <https://datareportal.com/reports/digital-2021-france>.
- [3] Vasquez-Henriquez P, Graells-Garrido E, Caro D. Tweets on the go: Gender differences in transport perception and its discussion on social media. Sustainability. 2020;12(13):5405. Publisher: Multidisciplinary Digital Publishing Institute.
- [4] Bode L, Budak C, Ladd JM, Newport F, Pasek J, Singh LO, et al. Words That Matter: How the News and Social Media Shaped the 2016 Presidential Campaign. Brookings Institution Press; 2020. Available from: <https://books.google.es/books?id=JbpkDgAAQBAJ>.
- [5] Ortega-Sánchez D, Blanch JP, Quintana JI, Cal ESdl, de la Fuente-Anuncibay R. Hate Speech, Emotions, and Gender Identities: A Study of Social Narratives on Twitter with Trainee Teachers. International Journal of Environmental Research and Public Health. 2021;18(8). Available from: <https://www.mdpi.com/1660-4601/18/8/4055>.

- [6] Wood-Doughty Z, Xu P, Liu X, Dredze M. Using noisy self-reports to predict twitter user demographics. arXiv preprint arXiv:200500635. 2020.
- [7] Yang YC, Al-Garadi MA, Love JS, Perrone J, Sarker A. Automatic gender detection in Twitter profiles for health-related cohort studies. JAMIA open. 2021;4(2):ooab042. Publisher: Oxford University Press.
- [8] Vicente M, Batista F, Carvalho JP. In: Kóczy LT, Medina-Moreno J, Ramírez-Poussa E, editors. Gender Detection of Twitter Users Based on Multiple Information Sources. Cham: Springer International Publishing; 2019. p. 39-54. Available from: https://doi.org/10.1007/978-3-030-01632-6_3.
- [9] Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, et al. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In: The World Wide Web Conference. WWW '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 2056-67. Available from: <https://doi.org/10.1145/3308558.3313684>.
- [10] Liu Y, Singh L, Mneimneh Z. A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In: Proceedings of the International Conference on Deep Learning Theory and Applications; 2021. .
- [11] Bianchi F, Cutrona V, Hovy D. Twitter-Demographer: A Flow-based Tool to Enrich Twitter Data. arXiv preprint arXiv:220110986. 2022.
- [12] Steinert-Threlkeld ZC. Twitter as Data. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press; 2018.
- [13] Ikae C, Savoy J. Gender identification on Twitter. Journal of the Association for Information Science and Technology. 2022;73(1):58-69. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24541>. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24541>.
- [14] Rangel F, Rosso P, Koppel M, Stamatatos E, Inches G. Overview of the author profiling task at PAN 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. CELCT; 2013. p. 352-65.
- [15] Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, et al. Overview of the 2nd author profiling task at pan 2014. In: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014; 2014. p. 1-30.
- [16] Rangel Pardo FM, Celli F, Rosso P, Potthast M, Stein B, Daelemans W. Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers; 2015. p. 1-8.
- [17] Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B. Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.; 2016. p. 750-84.
- [18] Rangel F, Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF. 2017:1613-0073.
- [19] Rangel F, Rosso P, Montes-y Gómez M, Potthast M, Stein B. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF. 2018:1-38.
- [20] Rangel F, Rosso P. Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in twitter. In: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop; 2019. .
- [21] Michael J. 40 000 Namen. Anredebestimmung anhand des Vornamens. c't. 2007 Aug;17:182-3. Place: Hannover Publisher: Heise Zeitschriften Verlag. Available from: <http://www.heise.de/ct/ftp/07/17/182/>.
- [22] Knowles R, Carroll J, Dredze M. Demographer: Extremely Simple Name Demographics. In: Proceedings of the First Workshop on NLP and Computational Social Science. Austin, Texas: Association for Computational Linguistics; 2016. p. 108-13. Available from: <https://aclanthology.org/W16-5614>.
- [23] Wood-Doughty Z, Andrews N, Marvin R, Dredze M. Predicting Twitter User Demographics from Names Alone. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. New Orleans, Louisiana, USA: Association for Computational Linguistics; 2018. p. 105-11. Available from: <https://aclanthology.org/W18-1114>.
- [24] Hu Y, Hu C, Tran T, Kasturi T, Joseph E, Gillingham M. What's in a name? – gender classification of names with character based machine learning models. Data Mining and Knowledge Discovery. 2021 Jul;35(4):1537-63. Available from: <https://doi.org/10.1007/s10618-021-00748-6>.
- [25] Devlin J, Chang MW, 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, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86. Available from: <https://aclanthology.org/N19-1423>.
- [26] Wood-Doughty Z, Mahajan P, Dredze M. Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. New Orleans, Louisiana, USA: Association for Computational Linguistics; 2018. p. 56-61. Available from: <https://aclanthology.org/W18-1108>.
- [27] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 38-45. Available from: <https://aclanthology.org/2020.emnlp-demos.6>.
- [28] Kirk H, Jun Y, Iqbal H, Benussi E, Volpin F, Dreyer FA, et al.. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv; 2021. Available from: <https://arxiv.org/abs/2102.04130>.
- [29] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. arXiv; 2015. Available from: <https://arxiv.org/abs/1506.02640>.
- [30] Sap M, Park G, Eichstaedt J, Kern M, Stillwell D, Kosinski M, et al. Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1146-51. Available from: <https://aclanthology.org/D14-1121>.
- [31] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997 nov;9(8):1735-1780. Available from: <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [32] Chung J, Cho K, Bengio Y. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016. p. 1693-703. Available from: <https://aclanthology.org/P16-1160>.
- [33] Faruqui M, Tsvetkov Y, Neubig G, Dyer C. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics; 2016. p. 634-43. Available from: <https://www.aclweb.org/anthology/N16-1077>.
- [34] Rizve MN, Duarte K, Rawat YS, Shah M. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. arXiv:210106329 [cs]. 2021 Apr. ArXiv: 2101.06329. Available from: <http://arxiv.org/abs/2101.06329>.

Analysing Food-Porn Images for Users' Engagement in the Food Business

V.CASALES-GARCIA^{a,1}, Z.FALOMIR^b, L.MUSEROS^b, I.SANZ^b, D.M.LLIDO^b
and L.GONZALEZ-ABRIL^a

^aUniversidad de Sevilla, Avda. Ramón y Cajal, 1, Sevilla, E-41018, Spain

^bUniversitat Jaume I, Av. Vicent Sos Baynat s/n, Castelló, E-12071, Spain

Abstract. This paper presents an approach for analysing *food-porn* images and their related comments published by the cooking school *Getcookingcanada* Instagram account. Our approach processes the published images to extract colour parameters, counts the number of *likes*, and also analyses the comments related to each publication. A dataset containing all these was built, and methods were applied to study correlations among the data: a regression analysis, an ANOVA and a sentiment analysis of the comments on the dataset to explain the relation between the quantity of likes and the sentiment obtained from the food images. Our results show a correlation between the number of *likes* and the sentiment analysis of the comments. Images that evoke a positive sentiment have a higher number of likes and comments. Users' experience on creating posts is also analysed and confirms a positive correlation between the number of *likes* and the publisher's experience.

Keywords. Sentimental analysis, Food-porn, regression, Anova, deep learning.

1. Introduction

Food-porn refers to how people share images of food through social media in order to have an impact on potential consumers. For social media publications (e.g. in Instagram) the number of *likes* associated with them are important since a higher number of *likes* involves a larger number of followers and, therefore, an increased impact. Hence, *food-porn* images is a way for small and medium enterprises (SMEs) to create loyal customers and promote gastronomic tourism.

In the literature, diverse works have also deal with the problem of analysing images and comments appearing in social media [1,2,3,4,5]. According to a study by [1], the emotion of Gastronomic Tourism Experiences on Digital Media Platforms took 25,000 photos of Instagram. The result shows that most gourmet tourism content is positively received across all platforms. And according to [2], Food Brands use social media platforms such as Instagram to market their products to a growing number of consumers, using a high frequency of targeted and curated posts that manipulate consumer emotions rather than present information about their products.

¹All authors contributed equally to this paper.

2. Methods

This study is based on a dataset of 1523 culinary images published in the Instagram account @getcookingcanada² by an online cooking school. These images and their associated meta-information were retrieved using the Instagram API. The relevant meta-information includes the textual description of the image, the quantity of *likes* and comments posted by the followers. Among the 1523 images, 958 had comments, with an average of 3.05 comments per image. The average quantity of *likes* for each image was approximately 47.

The *polarity* of the comments was obtained using a model based on DistilBERT [6], a deep learning model based on transformers. The score obtained was discretised converting it into the qualitative labels *Positive*, *Neutral* and *Negative*. Instagrammers' *experience* is measured as years pass by as their uploaded photos become more attractive to their followers progressively.

The *compression factor* of each food image was computed (ratio between the JPEG-compressed image at 100% quality and the full-size uncompressed image) and *colour-related metrics* were also obtained: the 5 predominant colours (obtained by a deep learning model to detect the food within the image [7], and then applying the median cut algorithm to the resulting, cropped image), the *colourfulness* (a linear combination of the mean and standard deviation of the pixel cloud in the colour plane), and the *number of distinct colours* (12-bit-quantised) in the RGB image. A *likelihood ranking for each colour palette* (from 1 to 5) was also computed based on a model obtained by learning from the ColorLovers dataset [8].

3. Results

Figure 1(a) shows how the @getcookingcanada account grows as the quantity of *likes* increases (followers reaffirm themselves on Instagram) as years pass by. Moreover, the quantity of comments increases in line with the increasing tendency of *likes*. Negative comments do not increase, but they tend to disappear. Let us indicate that if a food image has no comments then it is tagged as neutral. Figure 1(b) shows that negative items tend to go down, positive ones are increasing, and neutral tones, go down substantially, showing a stronger commitment by the followers to add comments to the posts.

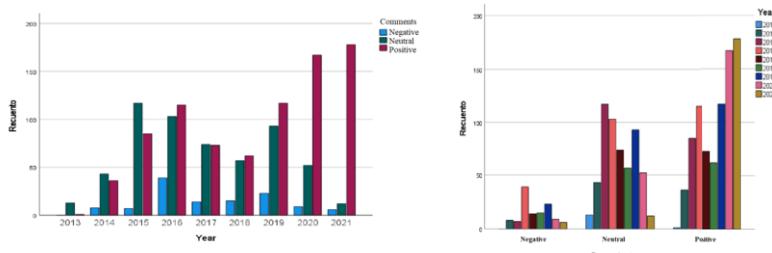


Figure 1. Analysis of followers' comments per year.

²Granted their permission to us for processing images. Note also that, our analysis only took into account the food images, other images containing people, such as a chef or student's, were discarded.

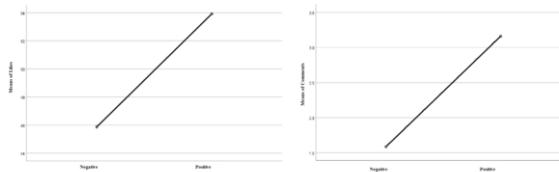


Figure 2. Comparisons of means regarding the likes and the negative/positive sentiments, and (b) Comparisons of means ANOVA regarding the number of comments and the negative/positive sentiments.

Two models are considered in order to find out the relation influencing the quantity of likes in the postings: Model 1 and ANOVA. It is worth noting that the dataset was pre-processed to remove outliers and that images without comments were removed.

Model 1. where the *likes*, denotes by y_i , is the dependent variable and the independent variables are: e_1 =experiences, c_2 =comments and sa_3 =sentimental analysis. That is, the model considered is: $y_i = \beta_0 + \beta_1 e_{i1} + \beta_2 c_{i2} + \beta_3 sa_{i3} + u_i$.

Table 1. (a) Summary of models (b) Coefficients. Dependent Variable: likes

Model	Adjusted Standard				Model	Coefficients			
	R	R^2	R^2	Error		no standards	standards	t	p-value
1	.710	.504	.503	15.238	(Constant)	20.650	0.898	- - - -	22.987 .00
2	.726	.527	.525	14.907	Experiences	0.013	0.001	0.474	24.306 .00
					Comments	3.202	0.182	0.385	17.597 .00
					Sentimental A.	-0.389	1.275	-0.006	-0.306 .76

Table 1 shows that the experiences and comments are significant but not the results on sentimental analysis (level of significance at 5%). The experience variable is more relevant than the comments variable since its coefficient standard is bigger (.474 > .385). Thus, if the experience and the comments increase, then the likes increase too. The sentimental analysis variable does not provide information to explain the likes. We hypothesised that if the comments are positive, this will result in followers giving a like. However, this is not corroborated in the model. We think that this is due to the fact that the tool for calculating the sentiment score was not able to collect the followers' sentiments. Model 1 explains the 50,4% of the variability of the *likes* of the followers.

ANOVA. Analyzes whether a positive or negative sentiment influence the likes and the number of comments given by the followers. With respect to the likes, Table 2 and Figure 2 (left) show that there is a significant difference between the average of the likes for positive and negative sentiments. When the followers have a favourable impression (positive sentiment) that is demonstrated by the followers participating more positively and contributing a more significant quantity of likes. On the contrary, there are fewer likes if the impression caused by the image is negative (negative sentiment). With respect to the number of comments, Table 3 and Figure 2 (right) show that when the followers have a

Table 2. ANOVA: Dependent variable: likes. Independent variables: + and - sentiments.

Model	Sum of squares	gl	Quadratic Mean	F	Sig.
Regression	6851.789	1	6851.789	13.807	.000
Residual	469960.091	947	496.262		
Total	476811.880	948			

favourable impression they tend to participate more positively and make a more significant number of comments. On the contrary, there are fewer comments if the impression caused by the image is negative.

Table 3. ANOVA: Dependent variable: number of comments. Independent variables: sentiments Negative and Positive.

Model	Sum of squares	gl	Quadratic Mean	F	Sig.
Regression	261.503	1	261.503	35.979	.000
Residual	6882.975	947	7.268		
Total	7144.478	948			

4. Conclusion

Culinary photos are crucial to promote the food business, and the images colour and texture are features that could evoke a response that might be positive or negative. According to our results the food images with less complex colours and texture evoked more positive responses or likes. This might be due to the food images are more aesthetic. Finally, the evoked positive emotion may produce a good response for consumers. We intend to carry out a survey in the nearer future in order to test these hypotheses.

We conclude that social networks can be a great promoter of food business and generate followers and loyal consumers to our products. Therefore, the increase in likes and comments from hashtags is suitable for generating greater diffusion of a product, so we see Instagram as a medium for promoting tourist destinations in general, not only for promoting online cooking schools as [@getcookingcanada](#).

Acknowledgments. This work has been funded by the projects FPU 17/00014, PDC2021-121097-I00, RYC2019-027177-I / AEI / 10.13039/501100011033, UJI-B2020-15, PGC2018-102145-B-C21. The authors acknowledge [@getcookingcanada](#) data and Chef Kathryn Joel.

References

- [1] K S G, Sinnoor G. Analysis of User-Generated Contents in Digital Media towards Gastronomic Tourism Experiences: Sentimental and Locational Approach. Turkish Online Journal of Qualitative Inquiry. 2021 08;12:4170-5.
- [2] Vassallo A, Kelly B, Zhang L, Wang Z, Young S, Freeman B. Junk Food Marketing on Instagram: Content Analysis. JMIR Public Health and Surveillance. 2018 06;4:e54.
- [3] Sharma SS, De Choudhury M. Measuring and Characterizing Nutritional Information of Food and Ingestion Content in Instagram. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery; 2015. p. 115–116.
- [4] Kusumasondjaja S, Tjiptono F. Endorsement and visual complexity in food advertising on Instagram. Internet Research. 2019 02;29.
- [5] Filieri R, Yen DA, Yu Q. #ILoveLondon: An exploration of the declaration of love towards a destination on Instagram. Tourism Management. 2021;85:104291.
- [6] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:191001108. 2019.
- [7] Park D, Lee J, Lee J, Lee K. Deep Learning based Food Instance Segmentation using Synthetic Data. In: 2021 18th International Conference on Ubiquitous Robots (UR). IEEE; 2021. p. 499-505.
- [8] Musersos L, Sanz I, Falomir Z, Gonzalez-Abril L. Creating, Interpreting and Rating Harmonic Colour Palettes Using a Cognitively Inspired Model. Cognitive Computation. 2020;12(2):442-59.

Influence in Social Networks Through Visual Analysis of Image Memes

Carles ONIELFA^{a,1}, Carles CASACUBERTA^a and Sergio ESCALERA^{a,b}

^aDepartament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

^bComputer Vision Center, Spain

Abstract. Memes evolve and mutate through their diffusion in social media. They have the potential to propagate ideas and, by extension, products. Many studies have focused on memes, but none so far, to our knowledge, on the users that post them, their relationships, and the reach of their influence. In this article, we define a meme influence graph together with suitable metrics to visualize and quantify influence between users who post memes, and we also describe a process to implement our definitions using a new approach to meme detection based on text-to-image area ratio and contrast. After applying our method to a set of users of the social media platform Instagram, we conclude that our metrics add information to already existing user characteristics.

Keywords. Memes, clustering, social media, social network, influence, culture, DBSCAN, CNN, graph

1. Introduction

A meme is usually defined as “an idea, behavior, phrase or usage that spreads within a culture” [1]. In the digital era, memes have adapted to new technologies and have become a phenomenon in contemporary web culture [2]. As a combination of humor, text, and a symbol, emoticons became one of the first types of Internet memes.

Even though Internet memes can exist as text, emojis, videos, or gifs, a common format is that of an image with superimposed text that conveys some type of merged message in an epigrammatic style. In the earlier days of the Internet, images with superimposed text began to propagate via e-mail and message boards. Later, social networks emerged, allowing memes to viralize [3]. Image memes have become an integral part of Internet culture. With the help of users they are born and reproduced, often mutated in the process. They are also used to spread political messages and ideologies. Compared to textual memes, image memes can condense their content and require less attention to be understood. Therefore, they are likely to be more effective [4].

Many studies have been carried out around memes, mainly focusing on their evolution [5], predicting their virality [4, 6], modeling their spread with mathematical models [7, 8], or devising algorithms for detecting them [3, 9]. But few, if any, have dug deeper behind the creators of memes.

¹Corresponding Author: Carles Onielfa, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain; carlesonielfa@gmail.com.

This work has been partially supported by MICIN/AEI under projects PID2019-105093GB-I00 and PID2020-117971GB-C22, and by ICREA under the ICREA Academia programme.

Regarding human achievement, viral success is closely related to merit [10]. Therefore, it is natural that memes that were once uploaded anonymously are now being uploaded by users that are proud of their creations and sign their memes with their watermark. Some users who post popular memes have achieved massive followings, and this grants them enormous influence and reach. However, that would be true of any user on a social network with a big number of followers. What makes meme creators unique is that they not only have the power to reach their followers, but two factors greatly expand their scope. First, memes are meant to be spread and shared; hence, followers of meme creators, if they enjoy a meme, are likely to share it with their friends [11]. Second—and most importantly—from an original meme, other creators can mutate and alter the original to make their own, retaining core aspects of the meme such as the underlying image. If there was an idea or product within it, as the meme and its mutations viralize and are shared, the idea or product goes viral with it, achieving exposure orders of magnitude greater than the original reach of the creator of the meme.

In this study, we take over the task of providing tools to gain insight into creators and the relationship between them through a visual analysis of the content of their memes. Specifically, we provide a definition of a graph for visualizing relationships between users who post memes on social networks, together with metrics that evaluate and rank users (Section 3) and a process for experimentally building the corresponding graph (Section 4). We undertake the detection of image memes using a new approach, namely the extraction of features using a convolutional neural network (CNN) and the clustering of memes by their underlying images. As an example, we apply our definitions to a set of Spanish users of Instagram who publish memes, and comment on the results (Section 5).

We conclude that, thanks to our ranking, we are able to determine the users with the biggest potential for publishing viral memes, and that some of these would be overlooked by using standard metrics for determining influence.

2. Related Work

Recent studies in network analysis have analyzed how culture and behavior are spread via social network ties, yet without focusing on the phenomenon that revolutionized culture spread in social networks, namely memes. Likewise, studies in computer vision have analyzed image memes and have attempted to detect memes or cluster them together, without taking a look at the users who post them. This study bridges the gap.

User Influence in Social Networks. Many studies have reported that behaviors or preferences of people can spread via social ties in social networks, mainly getting their knowledge through surveys [12, 13, 14]. However, to our knowledge, only [15] has derived the users' characteristics from what the users post online. In [15] a CNN was used to classify the images that a user posted online. Then, the categories of the images of the users were compared among friends and random users to find that socially tied individuals are more likely to post images showing similar cultural lifestyles.

Meme Detection and Clustering. There have been studies that use memes and phrases extracted from news and blogs to track and study the dynamics of the news cycle [16] and research into clustering text-based memes on Twitter [17]. More in line with our study, the research in [18] was able to cluster image streams using perceptual image hashes (pHash). They identify memetic clusters using meme annotation from sites such as “Know Your Meme”. One recent approach to meme detection is the Memesequencer

model developed in [9]. However, research in [9] is limited to memes that have identifiable templates, previously documented on sites like Memegenerator or Quickmeme. Another approach to meme detection is the Meme-Hunter model from [3], which uses multi-modal deep learning. Their model combines image features, text features, and facial detection. However, they only consider memes as pictures with superimposed text in impact font or text placed in white space over a picture.

In comparison with these works, our approach to meme detection fits a much broader definition of meme and is more in line with the ever-changing landscape of memes, as it does not require the template to be previously catalogued.

3. Formalization of the Problem

In this section, we detail concepts about memes and formalize the context of our problem.

Definition 1. A *meme* is a virally transmitted image embellished with text, usually sharing pointed commentary on cultural symbols, social ideas, or current events.

This definition of meme could be expanded to contain videos, text or simply cultural references. However, within this study we only consider image memes.

Examples of memes can be seen in Fig. 1. Given a meme, we refer to its *meme format* as the underlying image of the composition. A meme format can often be used to create more than one meme by adding or changing the existing embellishments. An example of a meme format is shown in Fig. 2.



Figure 1. Three memes.



Figure 2. A meme format, known as “Galaxy Brain”.

For any meme format, there exists a meme that used it first. Given a set of users U and a meme format F , the *pioneer* of F within U is the user $u \in U$ who published the oldest meme with the format F . If the set of users is the set U_{tot} of all users on all social media platforms, then we refer to the pioneer as an *absolute pioneer*.

Definition 2. Given a set of users $U = \{u_0, \dots, u_n\}$, not necessarily belonging to the same social network platform, we define the *meme influence graph* of U as a directed weighted graph (U, E, w) with the following properties:

1. A pair (u_i, u_j) with $i, j \in \{0, \dots, n\}$ and $i \neq j$ is in the set E of edges if the user u_j has posted a meme whose format was pioneered by u_i .
2. $w(u_i, u_j)$ is the number of memes posted by u_j whose format was pioneered by u_i .

A meme influence graph $M = (U, E, w)$ is called *maximal* if $U = U_{\text{tot}}$, that is, if every user in every social media platform is in the set U .

Definition 3. Let (U, E, w) be a meme influence graph for users $U = \{u_0, \dots, u_n\}$.

1. The *out-degree* of a user u_i is the number of outgoing edges $(u_i, u_j) \in E$ from u_i , that is, the number of other users who have used a meme format pioneered by u_i .
2. The *in-degree* of u_i is the number of incoming edges $(u_j, u_i) \in E$ to u_i , counting how many other users have pioneered meme formats that u_i has used.
3. The *weighted out-degree* of a user u_i is the sum $\sum_{j \neq i} w(u_i, u_j)$ of the weights of the outgoing edges from u_i . It is the number of memes published by other users who have used a meme format pioneered by u_i .
4. The *weighted in-degree* of a user u_i is the sum $\sum_{j \neq i} w(u_j, u_i)$ of the weights of the incoming edges to u_i , indicating how many memes have been published by u_i with a format pioneered by some other user in U .

The PageRank algorithm [19] applied to a graph measures the importance of each of its nodes taking into account the number of incoming edges and the importance of the source nodes of these edges. In short, a node will be important if other important nodes link to it. If A is the adjacency matrix for a graph (U, E, w) , the *reverse PageRank* of the node u_i is the value that the PageRank algorithm for the graph with adjacency matrix A^t (transpose of A) assigns to u_i . By computing the PageRank in this manner, one gives importance to the outgoing edges instead of the incoming edges.

Definition 4. The *score* of a user u_i is the value that the reverse PageRank algorithm assigns to u_i .

For a maximal influence graph, degrees can be interpreted as follows. The out-degree of u_i is the number of users who have been inspired by memes of u_i , while the in-degree is the number of users who have influenced u_i when creating memes. The weighted out-degree of u_i is the number of memes that have been influenced by u_i , and the weighted in-degree of u_i is the number of memes from u_i that have been influenced by some other user. Since a user, when creating a meme, can be inspired by a meme from a user who is not the pioneer of the meme format, the influence from a pioneer on a user is assumed to be indirect. In the case of a non-maximal influence graph, we can also use the previous interpretations but with some nuances. Suppose that the pioneer is not the absolute pioneer of a meme format. In that case, there might not even be an indirect relationship of influence, since given a user u_j in a set of users U who published a meme with a format F with pioneer $u_i \in U$, there exists a possibility that u_j first saw the format F from another user $u_k \notin U$. Therefore, the relationship of influence on a meme influence graph that is not maximal has to be interpreted as potential influence.

4. Implementation

The process for building a meme influence graph (Definition 2) is shown in Fig. 3. The input is a set of users U and the output is the meme influence graph for those users. Even though the meme influence graph is defined for users of any social media platform, in this study we limited the scope to Instagram. Since Instagram only allows users to publish images and videos, it is likely to find users whose content is mainly image memes. Furthermore, Instagram is the third biggest social media platform [20] and, on this platform, it is common for brands to partner with influential users (influencers) and publish sponsored posts [21]. Thus, metrics for determining influential users are valuable.

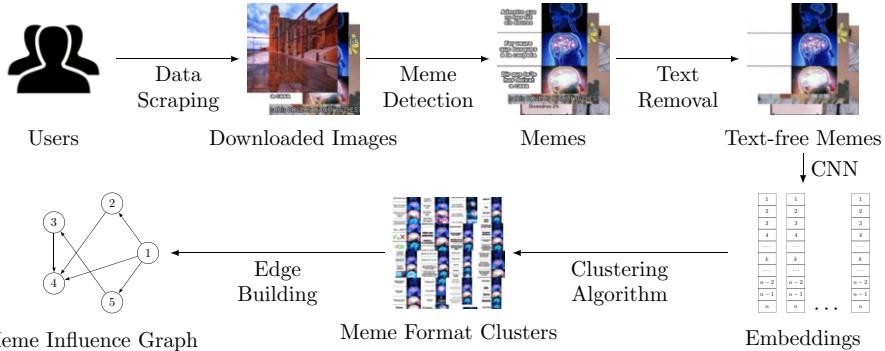


Figure 3. Flow diagram for the creation of a meme influence graph.

Data Scraping. Data for Instagram were extracted by storing the responses of Instagram’s API to the calls that the browser made when browsing relevant data. The data accessed in this study are 100% public and accessible by anyone. Retrieving user data is not strictly necessary for building a meme influence graph, but having some user characteristics enables us to interpret the graph and compare the metrics from the graph to the existing features of the users. Relevant features are their usernames, amount of followers, number of posts, average comments and likes per post, and text-to-image area ratio (computed after processing all the images from the user).

Algorithm 1 Meme Detection Algorithm

Require: $I :=$ image to process of size (w, h)
 $\alpha, \beta :=$ lower and upper bounds for the text-to-image area ratio
 $\gamma :=$ minimum standard deviation threshold

Ensure: *true* if the image is detected as a meme, *false* otherwise

```

1:  $p \leftarrow \text{text detection}(I)$             $\triangleright$  Detect areas containing text on the image using the CRAFT text detector [22]
2:  $A_{\text{tot}} \leftarrow w \times h$ ,  $A_{\text{text}} \leftarrow \text{area}(p)$            $\triangleright$  Compute total image area and text area
3:  $r \leftarrow A_{\text{text}}/A_{\text{tot}}$                    $\triangleright$  Compute text-to-image area ratio
4: if  $r \notin (\alpha, \beta)$  then                 $\triangleright$  If the text-to-image area ratio is not within bounds
5:   return false                          $\triangleright$  The image has either only text or no text and it is not a meme
6: end if
7:  $I_{\text{inpainted}} = \text{inpaint}(I, p)$      $\triangleright$  Inpaint  $I$  using the Navier–Stokes algorithm [23] with  $p$  as inpainting mask
8:  $\sigma = \text{std}(I_{\text{inpainted}})$              $\triangleright$  Compute standard deviation of grayscale values of the inpainted image
9: if  $\sigma \leq \gamma$  then                     $\triangleright$  If after text removal the image has high grayscale deviation
10:   return false                       $\triangleright$  There was no content of substance left after removing the text, so it is not a meme
11: else                                 $\triangleright$  If after text removal the image has low grayscale deviation
12:   return true                       $\triangleright$  There was an underlying image after removing the text, so the image is a meme
13: end if
```

When browsing the user’s publications (or posts), the only essential information are the images within them. As with the user data, features from publications were also valuable for later study, so they were extracted as well. Features extracted from posts were the user who posted it, the text, the amount of comments and likes, the date and time when the post was published, the hashtags added by the user, and whether the image had been detected as a meme or not (after processing it). In the case of Instagram, one publication can contain more than one media attachment (we call these publications *albums*) and the attachments can be images or videos. Videos have been treated as images using

their first frame. The first frame of a video is a good representation of the media in this context, since meme videos using the same format have very similar first frames.

Meme Detection. In line with our broad definition of meme, the task that the meme detection algorithm had to perform was to discard images with no text or no underlying image. The process used to perform this task is described in Algorithm 1.

Embeddings. From a text-free meme, we have to extract features to have a lower-dimensional representation of the source image that enables us to determine differences in content between two images by comparing their features. Using text-free memes instead of the original memes with text makes the underlying meme format exposed. This diminishes the differences between memes using the same meme format and makes it easier to cluster them together in the next step.

We use the convolutional neural network VGG16 [24] pre-trained with weights from the ImageNet challenge. This neural network was chosen because it gave good results for characterizing memes in [6], which had a broader meme definition than [3] and [9]. To adapt the network to the task at hand, we set the output to the second-to-last fully connected layer, bypassing the classifier layers and giving an output of 4096 dimensions.

Deep Image Clustering. To cluster memes into groups sharing the same meme format, we input the embeddings into a clustering algorithm, namely Density-Based Spatial Clustering of Applications with Noise (DBSCAN [25]). The DBSCAN algorithm works well with a large number of samples and uneven cluster sizes. It includes outlier removal while only requiring tuning of one parameter. We apply principal component analysis (PCA) to reduce the dimensionality of the samples to 1024 for improving efficiency.

Meme Influence Graph. Finally, we build the meme influence graph (Definition 2). We add input users as nodes and then, for each cluster, we create edges from the pioneer of a cluster to the authors of the rest of the memes of that cluster. After building the graph, we compute the metrics defined in Section 3.

Scalability. Meme detection and feature extraction are parallelizable. Clustering can be scaled by using an incremental DBSCAN implementation [26] or a highly parallelizable one [27]. Scaling is limited by the speed at which data can be extracted from Instagram.

5. Results

This section contains the results of using our implementation to build the meme influence graph for a selected set of users. The implementation was coded using Python. A No-SQL database was used for locally storing the data generated at each step of the process.

Data Scraping. The study set includes 91 users and 457,101 media. Users were selected starting from two sets of 5 and 11 users who posted memes in Spanish and Catalan, respectively, and had a large amount of followers. We added users who appeared in the “Related Accounts” section of the profiles of the starting users and also posted memes. This criterion was established to obtain a set of users that we expect to be densely connected in their meme influence graph. The amount of users was limited by the speed at which the data could be extracted from Instagram. The time frame of the posts was limited to a period comprised between January 1st, 2017 and April 23rd, 2022. Within the study set, there are users who post general topic memes but also some who post topic-specific memes, such as football-themed, music-themed, or region-themed.

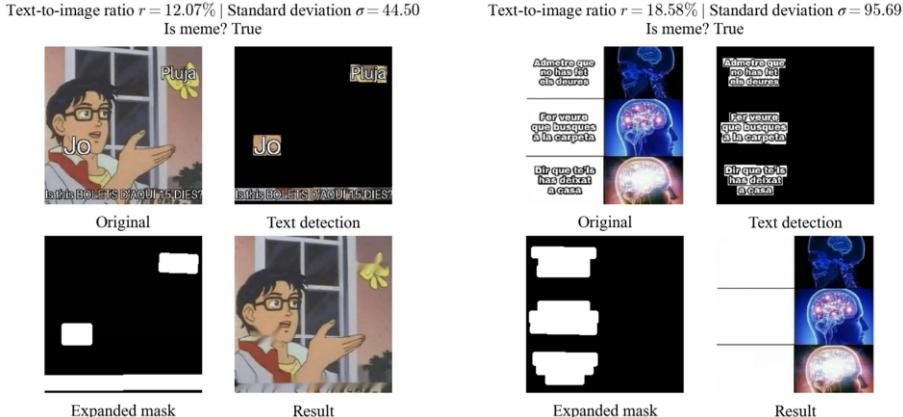


Figure 4. Images correctly detected as memes. Classification is based on text-to-image area ratio and standard deviation of gray values of the inpainted image converted to grayscale (Algorithm 1).

Meme Detection. The meme detection parameters (α, β, γ) in Algorithm 1 were found experimentally by selecting a random sample of the images on the dataset and splitting it into memes and non-memes using the meme detection algorithm. False positives and false negatives were manually identified and the thresholds were adjusted. This step was repeated several times until adjustments to the values were negligible. With this procedure, the following values were found: $\alpha = 0.018$, $\beta = 0.4$, $\gamma = 26$. Since standard deviation can vary depending on the size of the image, the images were resized to 224×224 pixels, matching the required dimensions for an input image to the neural network VGG16. In Figs. 4 and 5 we can see examples of how our meme detection algorithm processes the images. After applying the detection algorithm to all the images in our dataset, 342,984 out of 457,101 (75%) of images were detected as memes.

Embeddings and Clustering. The inputs to the VGG16 neural network are the text-free memes generated in the previous step, in a size of 224×224 pixels. The embeddings were reduced in dimensionality to 1024 components using PCA. The DBSCAN clustering algorithm was used with the cosine distance as metric, a minimum samples per cluster value of 3, and an epsilon value $\epsilon = 0.12$. The epsilon parameter for the algorithm was tuned manually by selecting a small number of memes with popular meme formats and visualizing their clusters with an initially big epsilon value. The epsilon value was lowered in small increments until the only memes left in the selected meme's cluster were memes with the same meme format. The clustering was able to group memes using the same meme format (Fig. 6). On our dataset, the algorithm found 13,663 clusters containing 82,801 memes, and 260,183 memes were detected as noise.

Influence Graph and Metrics. We built the meme influence graph and computed the metrics defined in Section 3. The graph for our anonymized set of users is shown in Fig. 7, where high score nodes can be easily identified. By computing the Pearson correlation coefficient between each of the pre-existing characteristics of the users and each of our metrics, we found that the follower count had low positive correlations with score ($\rho = 0.29$), weighted in-degree ($\rho = 0.27$), and weighted out-degree ($\rho = 0.30$). Hence we conclude that the number of followers, which is frequently used for determining the importance of a user [28], was unable to tell the difference between incoming influ-



Figure 5. Images correctly classified as non-memes because of high text area and low standard deviation of grayscale values, respectively (Algorithm 1).

ence and outgoing influence in our set of users. The amount of posts published by these users had a high correlation with score ($\rho = 0.69$), weighted in-degree ($\rho = 0.85$), and weighted out-degree ($\rho = 0.73$). This matches the intuition that the more posts a user makes, the more opportunities for their memes to influence or be influenced. No correlation higher than 0.10 was found for average likes and comments per post with influence per post (weighted out-degree/media count), indicating that user engagement in posts does not correlate significantly with more influential memes.

We found communities of users sharing memes that match a certain topic or geographic area using the Clauset–Newman–Moore community detection algorithm [29] on our graph. There were communities posting football-related memes and others related to territories. Most users were not included in any community with a relevant trait.

6. Conclusions

We presented a graph and metrics on it that serve as tools to visualize the influence and the relationships of meme creators, and provided a pipeline for constructing the graph and computing the metrics. This process was implemented using a novel approach to meme detection, deep features extraction, and DBSCAN clustering.

Our ranking method could be applied, for example, in order to select candidates from a set of users for a marketing campaign using memes. By basing our criteria on their

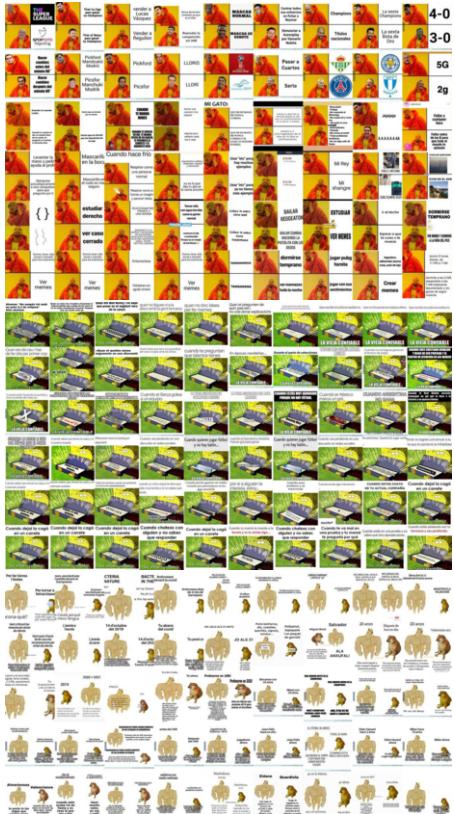


Figure 6. Some of the images in three clusters.

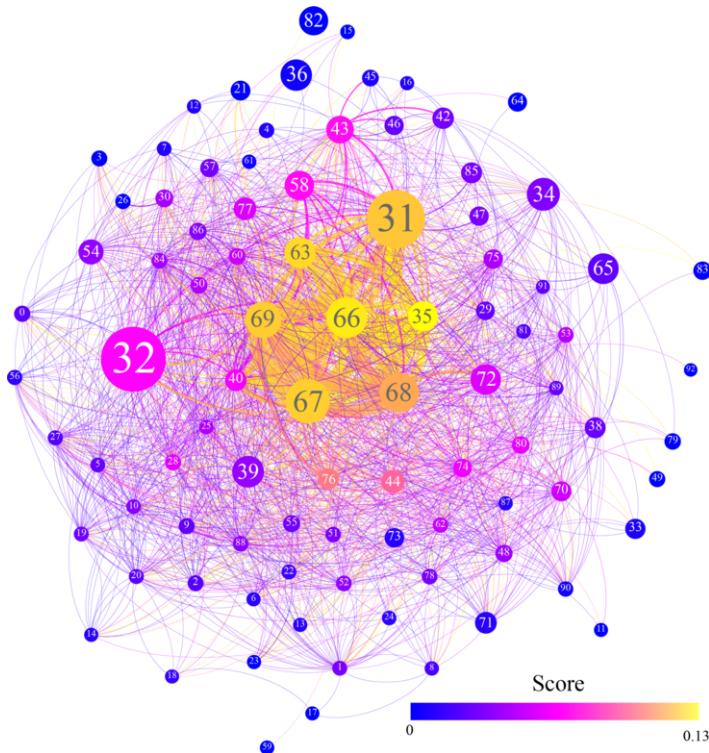


Figure 7. The meme influence graph for our set of 91 users. Nodes are labeled from 1 to 91; node size represents follower count; node color represents score from 0 (blue) to 0.13 (yellow); edge thickness represents edge weight; edge color matches the source node color; edge directions are represented clockwise.

scores, we ensure that memes generated by the selected users have the highest chance of viral spread through other users and reach an audience bigger than their group of initial followers. Using our graph, we can detect users who can be considered as “hidden gems”, that is, users with a high score although they may not rank high with respect to their number of followers. For example, user #35 in Fig. 7 has the highest score but ranks 13th regarding the number of followers.

Limitations. This small-scale experiment does not attempt to characterize Instagram as a social network or extract information about general meme format use or virality, although observing such characteristics would be feasible if the experiment were scaled to encompass enough users. The set of users in this article does not represent a general population; therefore, the methodology can be extrapolated to other sets of users but not the results. The metrics and connections also need to be carefully interpreted according to their definition as explained in Section 3, since it is very likely that users are related with other users not represented in the graph.

References

- [1] Blackmore S, Dugatkin LA, Boyd R, Richerson PJ, Plotkin H. The power of memes. *Scientific American*. 2000;283(4):64-73.

- [2] Laineste L, Fiadotava A. Globalisation and ethnic jokes: A new look on an old tradition in Belarus and Estonia. *The European Journal of Humour Research*. 2017;5(4).
- [3] Beskow DM, Kumar S, Carley KM. The evolution of political memes: Detecting and characterizing Internet memes with multi-modal deep learning. *Information Processing & Management*. 2020;57(2):102170.
- [4] Ling C, AbuHilal I, Blackburn J, De Cristofaro E, Zannettou S, Stringhini G. Dissecting the meme magic: Understanding indicators of virality in image memes. In: *Proceedings of the ACM on Human-Computer Interaction*. vol. 5. ACM New York, NY, USA; 2021. p. 1-24.
- [5] Bauckhage C. Insights into Internet memes. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 5; 2011. p. 42-9.
- [6] Barnes K, Riesenmy T, Trinh MD, Lleshi E, Balogh N, Molontay R. Dank or not? Analyzing and predicting the popularity of memes on Reddit. *Applied Network Science*. 2021;6(1):1-24.
- [7] Bauckhage C, Kersting K, Hadiji F. Mathematical models of fads explain the temporal dynamics of Internet memes. In: *AAAI Conference on Web and Social Media*. vol. 7; 2013. p. 22-30.
- [8] Weng L, Flammini A, Vespignani A, Menczer F. Competition among memes in a world with limited attention. *Scientific Reports*. 2012;2(1):1-9.
- [9] Dubey A, Moro E, Cebrian M, Rahwan I. Memesequencer: Sparse matching for embedding image macros. In: *Proceedings of the 2018 World Wide Web Conference*; 2018. p. 1225-35.
- [10] Yucesoy B, Barabási AL. Untangling performance from success. *EPJ Data Science*. 2016;5(1):1-10.
- [11] Wibowo T, et al. Usage of meme as information sharing media. *Jurnal Ilmiah Betrik: Besemah Teknologi Informasi dan Komputer*. 2020;11(3):165-71.
- [12] Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, et al. A 61-million-person experiment in social influence and political mobilization. *Nature*. 2012;489(7415):295-8.
- [13] Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*. 2008;30(4):330-42.
- [14] Christakis NA, Fowler JH. Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*. 2013;32(4):556-77.
- [15] You Q, García-García D, Paluri M, Luo J, Joo J. Cultural diffusion and trends in Facebook photographs. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 11; 2017. p. 347-56.
- [16] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle. In: *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2009. p. 497-506.
- [17] Ferrara E, JafariAsbagh M, Varol O, Qazvinian V, Menczer F, Flammini A. Clustering memes in social media. In: *Advances in Social Networks Analysis and Mining*. IEEE; 2013. p. 548-55.
- [18] Zannettou S, Caulfield T, Blackburn J, De Cristofaro E, Sirivianos M, Stringhini G, et al. On the origins of memes by means of fringe web communities. In: *Internet Measurement Conference*; 2018. p. 188-202.
- [19] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab; 1999.
- [20] Keepios. Global social media statistics; 2022. Available from: <https://web.archive.org/web/20220513111601/https://datareportal.com/social-media-users>.
- [21] Chen J. Instagram statistics you need to know for 2022; 2022. Available from: <https://web.archive.org/web/20220513234913/https://sproutsocial.com/insights/instagram-stats/>.
- [22] Baek Y, Lee B, Han D, Yun S, Lee H. Character region awareness for text detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2019. p. 9365-74.
- [23] Bertalmio M, Bertozzi AL, Sapiro G. Navier-Stokes, fluid dynamics, and image and video inpainting. In: *2001 IEEE Computer Society Conference on CVPR*. vol. 1. IEEE; 2001. p. I-I.
- [24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.
- [25] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *KDD'96, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; 1996. p. 226-31.
- [26] Chakraborty S, Nagwani NK. Analysis and study of incremental DBSCAN clustering algorithm. arXiv preprint arXiv:14064754. 2014.
- [27] Götz M, Bodenstein C, Riedel M. HPDBSCAN: highly parallel DBSCAN. In: *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*; 2015. p. 1-10.
- [28] Influencer Marketing: A Research Guide; 2020. Available from: <https://web.archive.org/web/20210627064440/https://guides.loc.gov/influencer-marketing/metrics-and-costs>.
- [29] Clauset A, Newman ME, Moore C. Finding community structure in very large networks. *Physical Review E*. 2004;70(6):066111.

Data Science, Recommender Systems and Decision Support Systems

This page intentionally left blank

Deep Air – A Smart City AI Synthetic Data Digital Twin Solving the Scalability Data Problems

Esteve ALMIRALL^{a,1}, Davide CALLEGARO^a, Peter BRUINS^a, Mar SANTAMARÍA^b, Pablo MARTÍNEZ^b and Ulises CORTÉS^c

^aEsade Business School, URL University

^b300.000km, 300000kms.net

^cKnowledge Engineering and Machine Learning Group, Universitat Politècnica de Catalunya

Abstract. Cities are becoming data-driven, re-engineering their processes to adapt to dynamically changing needs. A.I. brings new capabilities, effectively enlarging the space of policy interventions that can be explored and applied. Therefore, new tools are needed to augment our capacity to traverse this space and find adequate policy interventions. Digital twins are revealing themselves as powerful tools for policy experimentation and exploration, allowing faster and more complete explorations while avoiding costly interventions. However, they face some problems, among them data availability and model scalability. We introduce a digital twin framework based on an A.I. and a synthetic data model on NO₂ pollution as a proof-of-concept, showing that this approach is feasible for policy evaluation and (autonomous) intervention and solves the problems of data scarcity and model scalability while enabling city level Open Innovation.

Keywords. Digital Twins, Smart City Policy, Synthetic data, Digital twins and synthetic data

1. Introduction

Digital Twins have been evolving in aerospace exploration, manufacturing, and many other areas and, together with them, has been an evolution of their understanding.

Two challenges that Digital Twins find in cities are the lack of complete data, particularly real-time data, and the need for scalability. Cities are large, grow and change constantly and have fuzzy borders. The idea of sensorizing a whole city is undoubtedly bold, difficult to attain, challenging to make it economically sound, and even more to keep it updated.

In this paper, we introduce “Deep Air,” a proof-of-concept prototype to provide some light on solving these two problems using machine learning and synthetic data. In synthesis, our prototype is a digital twin for city pollution built with synthetic data cre-

¹Corresponding Author: Esteve Almirall, Esade Business School, URL University; E-mail: esteve.almirall@esade.edu

ated by a calibrated machine learning model. We show that the accuracy of the prototype is enough for investigating pollution city policies and reactions to specific conditions, taking into consideration the low granularity of applicable procedures [1].

This approach will not only help solve some of the problems of Smart City digital twins but also enables a model of decentralized self-regulated governance based on AI and synthetic data digital twins that we believe are endowed with higher agility, flexibility, scalability, and far lower cost while enabling Open, data-driven innovation in cities.

2. Materials and Methods

A digital twin is a virtual representation of the characteristics and behaviors of a physical entity used to study and predict its conduct without having to experiment with the actual object [2].

Smart Cities is probably today one of the most substantial areas of development of digital twins, not only in terms of projects being developed across the world in cities.

Smart City digital twins heavily depend on an abundance of data, particularly real-time data, with fine granularity.

The central idea of our proposal is to use synthetic data. Therefore, data is created through an intermediary A.I.-based model to feed the digital twin together with real data from sensors and synthetic data (see Fig 1).

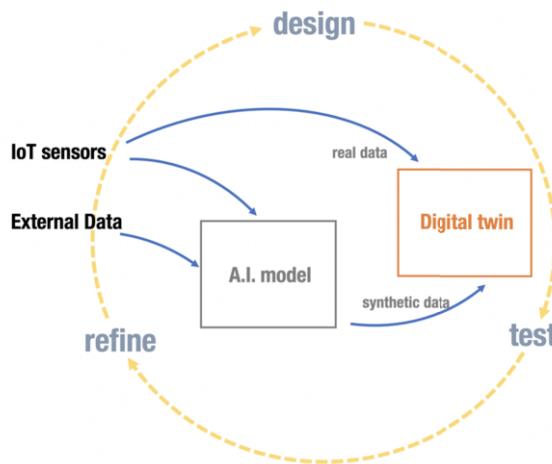


Figure 1. A digital twin fed with real and synthetic data in a design-test-refine loop.

The use of synthetic data will solve the problem of the scarcity of data. However, it could create another issue, the one of accuracy. We argue, however, that if the accuracy of the AI model supporting the digital twin is high enough and the granularity of the planning decision to be taken based on the data of the digital twin low enough, then there is a space where it will be no difference in terms of decisions between synthetic and real data. To express it formally, the equality of decision proposition must be satisfied.

Proposition 1. Equality of decisions.

Given a set of real data (measurements) $\Phi = \{\phi_1, \phi_2, \phi_3 \dots\}$ and a set of synthetic data (generated measurements) $\hat{\Phi} = \{\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3 \dots\}$ and given a digital twin $T(\Phi)$, where $T(\cdot)$ is a function assuming \mathbb{R}^∞ values, we can define the enabling decision function as $D(T)$ over a discrete space of decisions $D(T) = \{\delta_1, \delta_2, \dots\}$.

Thus, the equality of decisions proposition is satisfied when $D(T(\Phi)) = D(T(\hat{\Phi}))$, and this equality holds if and only if two conditions are met:

1. $\Phi \cong \hat{\Phi}$, therefore, a highly accurate model producing synthetic data is needed
2. $D = \{\delta_1, \delta_2, \delta_3, \dots\}$ met when δ_i is highly separated from $\delta_j, \forall_{i,j}$

In this paper, we will concentrate on the first condition 1), providing a proof-of-concept and showing that high accuracy (more than 88%) is possible in pretty complicated models (NO_2 congestion) with a minimal set of variables (eight in this case).

3. Results

Our results show that it is possible to predict NO_2 pollution data with an accuracy of 88.876%, std of 1.3768, with an XGBoost model primarily based on geographical data and only eight features. These are certainly encouraging results.

For this project, we have extracted data from multiple sources. Some data sets were publicly available, and others were given upon request.

- INE (Instituto Nacional de Estadística) data is all publicly available and found on their website [3].
- EEA (European Environmental Agency) data is publicly available and found on their website [4].
- The datasets held by 300000 km/s (300000kms.net) - an urban think tank located in Barcelona, which has been part of the research by providing insights and ad-hoc data - are private but not confidential and available upon request.

Initially, dataset collection resulted in approximately one hundred features. Using domain knowledge, an initial selection was made, resulting in a set of 28 elements that constituted our baseline model (Fig. 2).

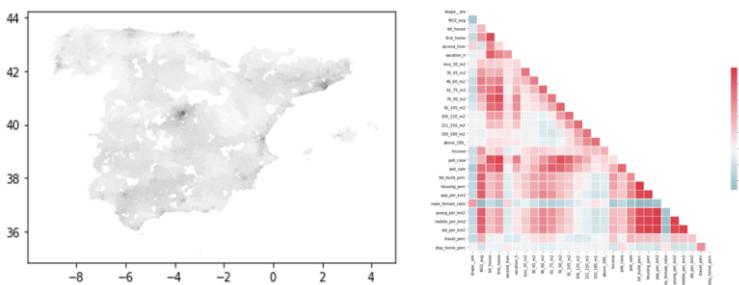


Figure 2. NO_2 data per district, together with the correlation matrix of the initial 28 features

Looking at feature importance, this model shows clear possibilities of simplification in addition to improvement using a more sophisticated algorithm.

We performed feature selection among the mix of lagged and non-lagged features, reducing them to only eight and using a Random Forest model.

Finally, we reproduce the same analysis with a more powerful model, an XGBoost, obtaining some improvement and reaching our final accuracy of 88.87%, std 1.376834 (See figure 3).

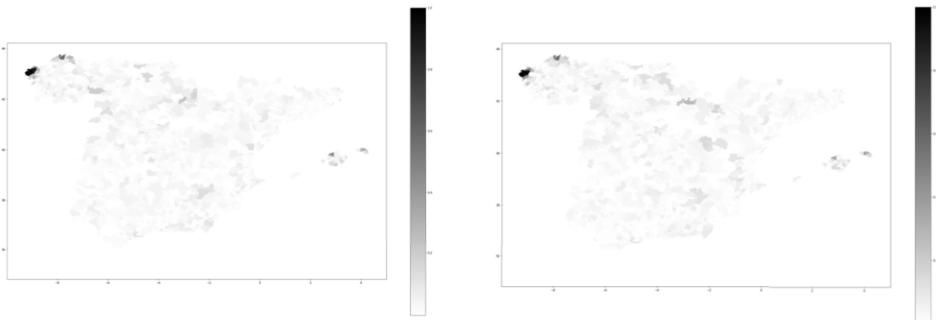


Figure 3. Random Forest and XGBoost model accuracy – darker is less accurate

As shown, only a few elements are determinants of NO₂ pollution, many of them static because they are part of the urban fabric. Therefore, it is possible to create models with limited features and high accuracy.

4. Conclusion

Through this paper, we have shown how synthetic data is feasible and solves many problems that today's digital twins face in complex socio-economic environments such as cities. Digital twins, in their broad sense as digital models that could accurately represent certain aspects of a city, allowing for experimentation, planning, knowledge discovery - metadata will be a valuable asset- and even near real-time adjustments or emergency activation procedures, are part of the future of management, optimization and city planning. Metaphorically we can say that code is the new concrete [5]. The proof of concept present in this work shows an alternative path that mixes real and synthetic data. We hope this research inspires a new generation of digital twins to support cheaper, more scalable, and open city management enabling open, data-driven innovation in cities.

References

- [1] Almirall, E and Callegaro, D and Bruins, P and Santamaría, M and Martínez, P and Cortés, U uh. Deep Air. A Smart City AI Synthetic data Digital Twin cracking the scalability data problems; 2022.
- [2] Jones D, Snider C, Nassehi A, Yon J, Hicks B. Characterising the Digital Twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology. 2020;29:36-52.
- [3] Instituto Nacional de Estadística. Census Data 2011; 2011. Available from: https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm.
- [4] European Environmental Agency. Air Pollution Data – EEA; 2019. Available from: <https://www.eea.europa.eu/themes/air/dc>.
- [5] LADOT. Los Angeles. Technology Action Plan; 2019. Available from: <https://ladot.io/wp-content/uploads/2019/03/LADOT-TAP-v7-1.pdf>.

Optimizing Online Time-Series Data Imputation Through Case-Based Reasoning

Josep Pascual-Pañach^{a,b,1}, Miquel Sànchez-Marrè^b, Miquel Àngel Cuguero-Escofet^c

^a Consorci Besòs Tordera, Catalonia (jpascual@besos-tordera.cat)

^b Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC)
Universitat Politècnica de Catalunya (UPC), Catalonia (miquel@cs.upc.edu).

^c Advanced Control Systems Research Group, Universitat Politècnica de Catalunya
(UPC), Catalonia (miquel.angel.cuguero@upc.edu).

Abstract. When working with Intelligent Decision Support Systems (IDSS), data quality could compromise decisions and therefore, an undesirable behaviour of the supported system. In this paper, a novel methodology for time-series online data imputation is proposed. A Case-Based Reasoning (CBR) system is used to provide such imputation approach. The CBR principle (i.e., solving the current problem using past solutions to similar problems) may be applied to data imputation, using values from similar past situations to replace incorrect or missing values. To improve the performance of the data imputation process, optimal case feature weights are obtained using genetic algorithms (GA). The proposed methodology is validated with data obtained from a real Waste Water Treatment Plant (WWTP) process.

Keywords. Online Data Imputation; Time-series; Case-Based Reasoning; Optimization; Intelligent Decision Support.

1. Introduction

Intelligent Decision Support Systems (IDSSs) operate using data obtained from different sources, such as sensors, and often in real time. The quality of these data is a common problem that should be tackled to ensure the good performance of the system. To solve the data imputation problem different machine learning techniques and models can be used. Here, a Case-Based Reasoning (CBR) approach is proposed in order to impute missing values in an online fashion, optimizing feature weights and considering the time through temporal CBR (TCBR). In [1] an imputation method based on a k nearest neighbours' algorithm is proposed and applied to a financial prediction problem. [2] proposes also the use of a reliable k nearest neighbours' (RKNN) algorithm applied to incomplete interval-valued data. In [3] a CBR approach for offline medium-gaps (from 3 to 10 missing values) imputation is proposed and applied to meteorological time series.

¹ Corresponding author: Josep Pascual-Pañach, Consorci Besòs Tordera, Av. Sant Julià, 241, 08403, Granollers, Catalonia; E-mail: jpascual@besos-tordera.cat.

2. Methods

In this section, we present a methodology to calibrate a CBR system to impute missing values from online time-series. Figure 1 shows how the imputation process through CBR is integrated in the classical CBR cycle.

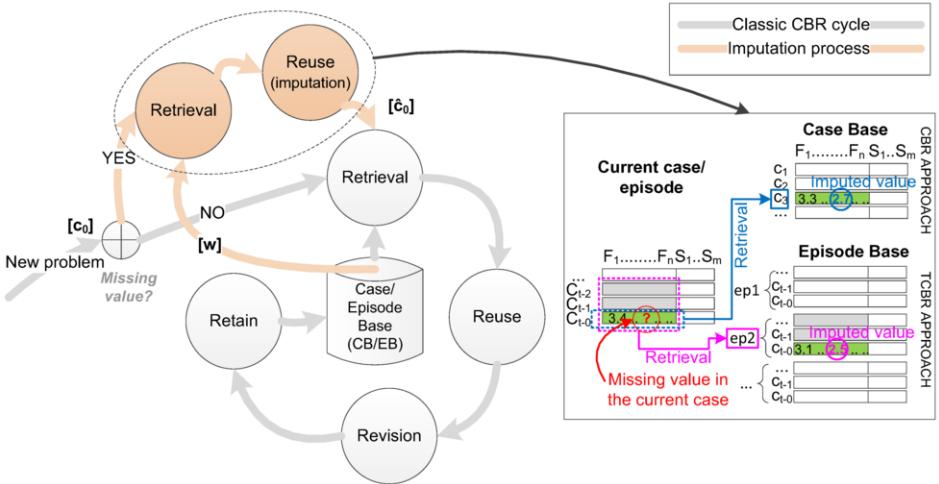


Figure 1. Integration of the CBR-based imputation approach in the CBR cycle

When a value from the current case c_0 is incorrect or missing, the available part of the case is used to find the most similar ones in the CB. Considering the TCBR approach, sets of consecutive cases (i.e., episodes) are used instead of particular cases in order to take into account data dynamics. This temporal approach is based on the one described in [4], and explained in another work under revision. The retrieval is done using episodes and giving the same importance to all cases. Regarding the imputation (reuse stage), in both CBR and TCBR the procedure is the same. The value of the missing feature is imputed using the corresponding value in the retrieved case or episode, obtaining a new case \hat{c}_0 . When using episodes, the value from the most recent case in the episode is used (which is the one corresponding to the most recent one in the current episode).

Assuming that all features are numeric, a weighted Euclidean Distance (wED) similarity measure is used in the retrieval stage as in Eq. (1).

$$wED(c_0, c_i) = \sqrt{\sum_{n=1}^N w_n (f_{0n} - f_{in})^2} \quad (1)$$

where c_0 and c_i are two cases, f_{0n} and f_{in} are the feature n values for each case, w_n is the weight for feature n and N is the number of features. Feature weights are calculated in order to minimize the error between the predicted value and the measured value for a particular feature. The metric used is the Root Mean Square Error (RMSE), calculated as in Eq. (2).

$$RMSE = \sqrt{\frac{1}{J} \cdot \sum_{i=1}^J (y(i) - \hat{y}(i))^2} \quad (2)$$

where J is the number of samples in the dataset, y is the measured data for a particular feature and \hat{y} is the predicted value for the same feature.

3. Experimentation

To evaluate the viability of the proposed imputation method, historical data from the real process under study has been used to generate the CB and to calibrate the imputation system by calculating an optimal vector w of feature weights. The optimization problem described in Section 2 has been solved using a Genetic Algorithm (GA). The problem to solve can be described as in Eq. (3):

$$\begin{aligned} \min_w e(w) \text{ subject to:} \\ \text{linear constraints: } [1 \dots 1] \cdot w = 1 \\ \text{bounds: } 0 \leq w \leq 1 \end{aligned} \quad (3)$$

where w is the weights vector and $e(w)$ is the cost function to be minimized. The cost function integrates Eqs. (1) and (2) to optimize the retrieval process with the aim of minimizing the RMSE between the measured and predicted values.

The method is integrated in an IDSS based on the integration of CBR and Rule-Based Reasoning systems used to set adequate operational set-points to control the biological process in a real Waste Water Treatment Plant (WWTP) [5]. Presented results correspond to the imputation of medium gaps –6 samples (30 minutes) and 12 samples (1 hour) of missing values due to typical real faults, e.g., communication faults or invalid values during the sensor calibration process (a common sensor maintenance periodic procedure in the real facility). Here, faults are simulated in order to have the measured values of the whole dataset to evaluate the performance of the method. Figure 2 shows the performance attained with both CBR approaches using unfaulty data.

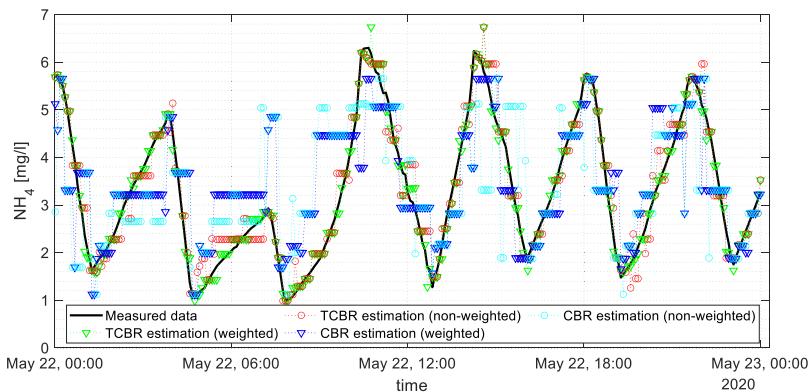


Figure 2. Different models are compared with unfaulty data

The best model (weighted TCBR) is validated with faulty data in Figure 3. Episodes have a fixed length of 6 samples, achieving a good trade-off between performance and computing time.

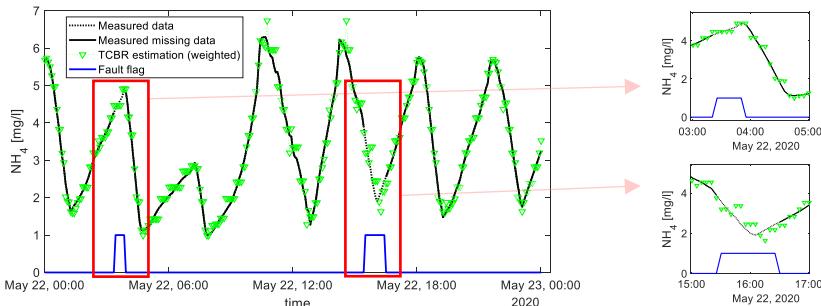


Figure 3. TCBR weighted model validated with 2 faults

4. Discussion, conclusions and next steps

This paper presents a data imputation method based on a CBR approach. The proposal has been evaluated using real data from a WWTP and considering different realistic medium missing data windows based on real faults in the ammonia sensor, which is a critical variable for the biological process control. An improved performance is obtained when using a calibrated CBR imputation system in comparison with the non-calibrated counterpart. The RMSE of the estimation with weighted features is almost 40% lower than the non-weighted estimation when using TCBR. Regarding the comparison between CBR and TCBR, the TCBR approach provides clearly better performance, with a RMSE about 60% lower than the calibrated CBR approach.

Next steps will consider a more in-depth evaluation of the method's performance with other sensors, different types of faults or multiple missing values, and the comparison with other classical time-series models and machine learning methods [6, 7] or state-of-the-art imputation techniques. Episodes' length will be also addressed.

References

- [1] Cheng, C.-H., Chan, C.-P., Sheu, Y.-J., 2019. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Engineering Applications of Artificial Intelligence* 81, 283–299.
- [2] Qi, X., Guo, H., Wang, W., 2021. A reliable KNN filling approach for incomplete interval-valued data. *Engineering Applications of Artificial Intelligence* 100, 104175.
- [3] Flores, A., Tito, H., Silva, C. (2019). CBRm: Case based Reasoning Approach for Imputation of Medium Gaps. *International Journal of Advanced Computer Science and Applications*. 10. 10.14569/IJACSA.2019.0100949.
- [4] Sánchez-Marrè, Miquel, Cortés, Ulises, Martínez, Montse, Comas, Joaquim, Rodríguez-Roda, Ignasi. (2005). An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*. 3620. 465-476. 10.1007/11536406_36.
- [5] Pascual-Pañach, Josep, Cugueró-Escofet, Miquel Àngel, Sánchez-Marrè, Miquel, Interoperating data-driven and model-driven techniques for the automated development of intelligent environmental decision support systems, *Environmental Modelling & Software*, Volume 140, 2021, 105021, ISSN 1364-8152.
- [6] Cugueró-Escofet, M. A., García, D., Quevedo, J., Puig, V., Espin, S., Roquet, J., 2016. A methodology and a software tool for sensor data validation/reconstruction: application to the Catalonia regional water network. *Control Engineering Practice* 49, 159–172.
- [7] Pascual-Pañach, J., Sánchez-Marrè, M., Cugueró-Escofet, M.A. (2022). *Ensemble model-based method for time series sensors' data validation and imputation applied to a real Waste Water Treatment Plant*. *11th International Congress on Environmental Modelling and Software, Brussels, Belgium*.

Synthetic Data for Anonymization in Secure Data Spaces for Federated Learning

Cecilio ANGULO^{a,1} Cristóbal RAYA^a

^a Universitat Politècnica de Catalunya. IDEAI-UPC, Spain

Abstract. Federated learning implies the integration of shared data. Privacy-enforcing platforms should be implemented to provide a secure environment for federated learning. We are proposing the integration of real world data from local data lakes and the generation and use of general synthetic data to simplify, eventually avoid, encryption or differential learning and use general architectures for data spaces.

Keywords. federated learning, synthetic data, data spaces

1. Introduction

Despite a large number of rich datasets are gathered across Europe that would be invaluable in the creation of knowledge through novel AI tools, the obstacles for shared use of these data hubs for data-driven innovation are insurmountable [1,2].

Federated Learning (FL) has recently emerged as a disruptive privacy technology allowing data use for model learning or data visualization, without exporting the data across from the data owner's hub [3,4]. This approach is being adopted as a key potential solution to the shared data use challenge. However, the FL approach is not mainly based on privacy protection, hence weakness issues has been pointed out for the general FL approach. Potential vulnerabilities include susceptibility to the man-in-the-middle attack and inference attacks aiming to re-identify data subjects [5].

Enhanced privacy methods being developed for federated learning includes processes such as Homomorphic Encryption (HE) [6], Differential Privacy (DP) [7] and Trusted Execution Environments (TEE) [8]. In particular, the latter one relates with the new directive about secure data spaces being promoted from the European Union (Data Governance Act).

In general, the term “data space” refers to a type of data relationship between trusted partners, each of whom apply the same high standards and rules to the storage and sharing of their data. Data are not stored centrally but at source and are therefore only shared (via semantic interoperability) when necessary. In this context, EU promoted initiatives

¹Corresponding Author: Cecilio Angulo, IDEAI-UPC, Jordi Girona 31, 08034 Barcelona, Spain; E-mail: cecilio.angulo@upc.edu.

as Gaia-X look for a trusted execution environment in the form of a secure data space where data is held exclusively by the members of the Association [9].

Data space's participants can be data providers, users and intermediaries. Data sovereignty and trust are essential for the working of secure data spaces and the relationships between participants. In this sense, references architectures like IDSA, from the International Data Spaces Association (IDSA), a founding member of the Gaia-X AISBL, has been proposed.

Unfortunately, these resource-intensive privacy enhancement methods, in the form of local software processes, developed hardware architecture or safe communication, severely limit the scalability of federated learning in the form of secure data spaces. Moreover, beyond privacy concerns or resources limitations, some critical data solutions imply storing and processing personal data to infer knowledge and to know about user experience, leading to legal consequences claiming.

Hence, anonymization arises as a keypoint in personal data manipulation [10,11], a tool to mitigate risks when gathering and massively processing sensitive data. This process allows identifying and shadowing sensitive information contained in documents, allowing its disclosure, hence avoiding to violate data protection rights of people and organisations that can be referenced in them [12].

Data anonymization, in other side, as a method protecting sensitive information or the identity of the data owner, due to legal or ethical issues, is usually seen as a major problem in data analytic because it could lead to reduce so much the knowledge contained into the dataset.

2. Proposal

It is worth noting that obtaining data has a high cost in so different domains and, many times, information is very limited. Therefore, many research projects have worked on developing reliable methods for data augmentation with synthetic instances. Moreover, knowledge extraction implies an experience acquired through learning, detecting patterns, looking for behaviours, assessing risks, until reaching a diagnosis and being able to propose a solution indicated for each situation. However, often, experts are unable to fully consider the large amount of data obtained from several institutions or companies and use it to make decisions. By considering the total set of data hubs, even for a focused problem, and generating synthetic experiences from a seedbed emerged from real-world data, professionals can benefit from this valuable information better than buried within huge amounts of data.

New training generative procedures, such as generative adversarial networks (GANs), aim at learning representations that preserve the relevant part of the information (about regular labels) while dismissing information about the private labels which correspond to the identity of a person. The success of this approach has been demonstrated in [13], for instance. As a result of the GAN-based anonymization phase, a seedbed is obtained from the training data that allows not only to capture information from the original data avoiding privacy concerns, but to generate new synthetic information with a similar behaviour to the original one. This result is currently being applied in generative applications on speech [14], vision [15], natural language [16] or in the health domain [17], where data are scarce and missing values are everywhere.

Our proposal advocates for the use of synthetic data for anonymization in the framework of secure data spaces with the aim of produce federated learning.

Let's us suppose N data hubs are storing information \mathcal{X} in the same domain. Currently, since data cannot be shared, models are being developed in local, $f_i = f(\mathcal{X}_i)$, for $i = 1, \dots, N$. In order to share information and federate learning, models are exported to other data hubs, evaluated on local data there, $f_i(\mathcal{X}_j)$ and fine tuned $\tilde{f}_{j,i}$ under the hypothesis that $\tilde{f}_{j,i} \sim \tilde{f}_{i,j} \sim f_{(i,j)} = f(\mathcal{X}_i \cup \mathcal{X}_j)$. However, it usually does not work. The claim for secure data spaces is that you can effectively share this information because the connection is safe, secure, and reliable. Hence, under the assumption that nobody will access your data, because either, you are moving models, not data, or data is moved in secure form, federated learning iterates so that you obtain $f_{(1, \dots, N)}$ from, for instance, $f_{1, \dots, N}$. In fact, there exist several approaches, but all of them have relaxed trust conditions.

Our approach is defending that you can share data, but not the original one, but one that is synthetically following the same statistical properties, that is $\hat{\mathcal{X}} \sim \mathcal{X}$ because $P(\mathcal{X}) = P(\hat{\mathcal{X}})$ with P indicating statistical properties. In this form, our hypothesis is that you can obtain $f_{1, \dots, N} = f(\bigcup \mathcal{X}_j)$ by federating $f_i(\mathcal{X}_i, \bigcup_{j \neq i} \hat{\mathcal{X}}_j)$ for $i = 1, \dots, N$.

In fact, as far as our approach is considering statistical properties, it can be noted from this domain in the form $P(\mathbf{y} | \bigcup \mathcal{X}_j) \sim \prod_i P(\mathbf{y} | \mathcal{X}_i, \bigcup_{j \neq i} \hat{\mathcal{X}}_j)$. This hypothesis is opening a new research line that we are starting to explore avoiding main privacy concerns.

References

- [1] Rahman S, Omar A, Bhuiyan M, Basu A, Kiyomoto S, Wang G. Accountable Cross-Border Data Sharing Using Blockchain Under Relaxed Trust Assumption. *IEEE Transactions on Engineering Management*. 2020 01;PP:1-11.
- [2] Gavrilov G, Vlahu-Gjorgjevska E, Trajkovik V. Healthcare data warehouse system supporting cross-border interoperability. *Health Informatics Journal*. 2020;26(2):1321-32. PMID: 31581924. Available from: <https://doi.org/10.1177/1460458219876793>.
- [3] Mohri M, Sivek G, Suresh AT. Agnostic Federated Learning. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 4615-25. Available from: <https://proceedings.mlr.press/v97/mohri19a.html>.
- [4] Bonawitz KA, Eichner H, Grieskamp W, Huba D, Ingberman A, Ivanov V, et al. Towards Federated Learning at Scale: System Design. In: SysML 2019; 2019. Available from: <https://arxiv.org/abs/1902.01046>.
- [5] Gambi S, Killijian MO, Núñez del Prado Cortez M. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*. 2014;80(8):1597-614. Special Issue on Theory and Applications in Parallel and Distributed Computing Systems. Available from: <https://www.sciencedirect.com/science/article/pii/S0022000014000683>.
- [6] Gentry C, Sahai A, Waters B. Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based. In: Canetti R, Garay J, editors. *Advances in Cryptology – CRYPTO 2013*. vol. 8042 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 75-92. Available from: http://dx.doi.org/10.1007/978-3-642-40041-4_5.
- [7] Torra V, Navarro-Arribas G. Integral Privacy. In: Foresti S, Persiano G, editors. *Cryptology and Network Security*. Cham: Springer International Publishing; 2016. p. 661-9.
- [8] Arfaoui G, Gharout S, Traoré J. Trusted Execution Environments: A Look under the Hood. In: Mobile-Cloud. IEEE Computer Society; 2014. p. 259-66. Available from: <http://dblp.uni-trier.de/db/conf/mobilecloud/mobilecloud2014.html#ArfaouiGT14>.

- [9] Dietrich M, Ferrari T. Governance, Architectures and Business Models for Data and Cloud Federations: the EOSC and GAIA-X Case Studies. Zenodo; 2021. Available from: <https://doi.org/10.5281/zenodo.4929021>.
- [10] Bae H, Jung D, Yoon S. AnomiGAN: Generative adversarial networks for anonymizing private medical data; 2019.
- [11] Bruynseels K, Santoni de Sio F, van den Hoven J. Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *Frontiers in Genetics*. 2018;9:31. Available from: <https://www.frontiersin.org/article/10.3389/fgene.2018.00031>.
- [12] Angulo C, Gonzalez-Abril L, Raya C, Ortega JA. A Proposal to Evolving Towards Digital Twins in Healthcare. In: Rojas I, Valenzuela O, Rojas F, Herrera LJ, Ortúñoz F, editors. *Bioinformatics and Biomedical Engineering*. Cham: Springer International Publishing; 2020. p. 418-26.
- [13] Piacentino E, Angulo C. Anonymizing Personal Images Using Generative Adversarial Networks. In: Rojas I, Valenzuela O, Rojas F, Herrera LJ, Ortúñoz F, editors. *Bioinformatics and Biomedical Engineering*. Cham: Springer International Publishing; 2020. p. 395-405.
- [14] Phan H, McLoughlin IV, Pham L, Chen OY, Koch P, De Vos M, et al. Improving GANs for Speech Enhancement. *IEEE Signal Processing Letters*. 2020;27:1700–1704. Available from: <http://dx.doi.org/10.1109/LSP.2020.3025020>.
- [15] Wang Z, She Q, Ward TE. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy; 2020.
- [16] Zhu Y, Zhang Y, Yang H, Wang F. GANCoder: An Automatic Natural Language-to-Programming Language Translation Approach based on GAN; 2019.
- [17] Piacentino E, Angulo C. Generating Fake Data Using GANs for Anonymizing Healthcare Data. In: Rojas I, Valenzuela O, Rojas F, Herrera LJ, Ortúñoz F, editors. *Bioinformatics and Biomedical Engineering*. Cham: Springer International Publishing; 2020. p. 406-17.

Towards Automated Compliance Checking of Building Regulations: smartNorms4BIM

Ignacio Huitzil ^{a,b,1}, Marco Schorlemmer ^a, Nardine Osman ^a, Pere Garcia ^a,
Josep Coll ^b and Xavier Coll ^b

^a Artificial Intelligence Research Institute (IIIA), CSIC, Bellaterra (Barcelona), Spain

^b Enginyeria i Project Management (EiPM), Barcelona, Spain

Abstract. This paper describes a preliminary approach towards automating the compliance checking of constructions with respect to building regulations. We describe a prototype that supports such automated checking by specifying regulations in terms of an ontology, and reasoning with the Building Information Models (BIM) of constructions. The first step in our approach is to translate regulations into a machine-readable format with the support of controlled natural language specifications of rules. Then, we propose a formal specification of the building regulations in OWL2, the de facto standard for ontology engineering on the web. We subsequently populate this ontology with data of real-world BIM specifications based on Industry Foundation Classes (IFC) in order to check their compliance with the formalized regulations. Finally, our prototype offers to the end-users a verification report in text and a graphical visualiser with the results of the compliance check. To explain how our prototype works and to demonstrate its applicability, we show some examples taken from a concrete use case.

Keywords. Building Information Modeling, Building Regulations, Ontologies, Rule-Compliance Checking

1. Introduction

In the field of architecture, engineering and construction (AEC), more and more standards are being used for the digitized representation of the physical and functional characteristics of a building. The regulations that affect this sector, however, are usually expressed in natural language and published in the official bulletins of local, regional, national and international governmental bodies. Therefore, the verification that the design of a building actually conforms to a certain regulation continues to be an intrinsically manual process, subject to human errors of interpretation, and it requires the experienced consultation of extensive documentation and data related to the construction.

In this paper, we describe SMARTNORMS4BIM, a prototype tool by means of which we attempt to automate part of the compliance checking process of building models as developed according to Building Information Modeling (BIM) standards [1], with

¹Corresponding Author, E-mail: ihuitzil@iiia.csic.es.

respect to construction norms and regulations. Any progress in the automation of this process can lead to a significant reduction in time, cost and risk of error. We show a proof of concept of our prototype tool for the *Decret d'Habitabilitat* of the Government of Catalonia [2] with respect to a particular BIM model of a building, carried out under the supervision of *Enginyeria i Project Management* (EiPM), an SME that has an extensive expertise in BIM project management.

We proceed as follows: Section 2 describes our proposal. Next, Section 3 discusses the main challenges we have encountered and the solutions we have adopted. In both sections, we show some illustrative examples. Then, Section 4 compares our approach with other related work, and, finally, Section 5 we describe our conclusions and future work.

2. Approach

We use the structure proposed by Eastman [3], which considers four stages for a rule-checking process : 1) translation of rules and regulations into a formal language; 2) preparation of the building model; 3) execution of the rule-checking process; and 4) reporting back the checking results. Figure 1 shows the architecture of our proposal.

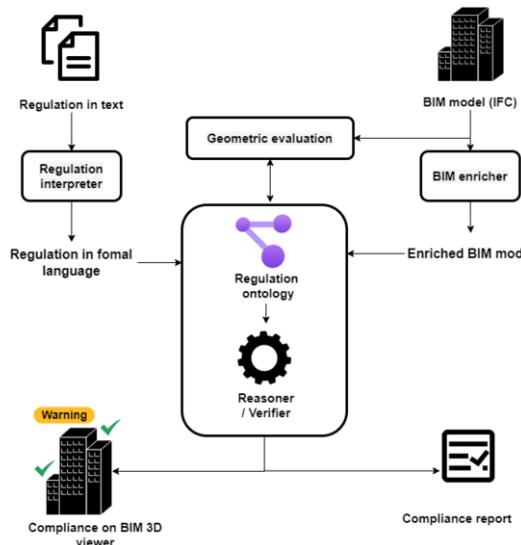


Figure 1. Architecture of the prototype.

2.1. Translation of rules and regulations into a formal language

For our approach we focused on Catalonia's building regulation and started by analyzing the “Decret 141/2012 sobre condicions mínimes d’habitabilitat dels habitatges i la cèdula d’habitabilitat” (from now on *Decret*), which specifies the minimal habitability

requirements for dwellings [2]. Originally in Catalan, we translated it into English (our working language) and reviewed Annex 1 and Annex 2, resulting in a total of 51 rules.

Our approach rests on the fundamental idea to take rules and regulations as defining relevant ontological classes and to see rule-checking as the reasoning task that attempts to classify the building elements of a particular BIM model according to these ontological classes. Compliance of a particular building with the norm will be given by how the reasoner classifies particular building elements to their respective expected classes. Let's take the following example of a couple of rules taken from the *Decret*:

Example 1.

Annex 1, Rule 3.11.2.² Hygienic appliances will be placed in bathrooms and their grouping is free (...).

Annex 1, Rule 3.15.³ Equipment. All dwellings must have: (...) b) A hygienic equipment that consists of at least one sink, one toilet and one shower.

We specify the rules in terms of ontological entities, namely by means of classes (e.g., *bathroom*, *shower*, etc.), and relations (e.g., a space, such as a bathroom, *having an equipment*, such as a shower). Furthermore, we take rules as defining our classes, in terms of necessary and sufficient conditions. For example, for Rule 3.15 we will define a class *3.15_validBathroom* to be a bathroom that has as equipment at least one sink, one toilet, and one shower. Example 2 shows this definition in a description logic.

Example 2.

$$\begin{aligned} 3.15_validBathroom &\equiv \text{Bathroom} \\ &\sqcap \exists \text{hasEquipment}. \text{Toilet} \\ &\sqcap \exists \text{hasEquipment}. \text{Sink} \\ &\sqcap \exists \text{hasEquipment}. \text{Shower} \end{aligned}$$

Figure 2 shows the same definition for valid bathrooms, expressed in Manchester OWL Syntax, in the “Equivalent To” area of the Protégé ontology editor.

For each building entity to which some rule applies (e.g., a bathroom) we define three classes: one for *valid* entities, namely those that comply with the regulation, one for *invalid* entities, namely those that violate the regulation, and one for entities that *lack data* to be considered either valid or invalid.

The collection of all formal definitions of the *Decret*'s rules constitute an ontology with respect to which we attempt to classify the actual building elements of a particular BIM model. This classification process captures thus the compliance checking process of the BIM model with respect to the *Decret*.

²“Els aparells destinats a la higiene se situaran a les cambres higièniques i la seva agrupació és lliure (...).”

³“Dotació/equip. Tots els habitatges han de disposar de: (...) b) Un equip higiènic que estigui format, com a mínim, per un rentamans, un vàter i una dutxa.”

2.2. Preparation of the building model

The BIM model we studied consists of one IFC file,⁴ (in IFC2x3 schema, file size 11.2 MB). This IFC file specifies one building of four stories with seven dwellings. They are located in the following manner, two of them on the first floor and the rest (five) on the second.

We took the BIM model expressed in IFC and extracted the relevant instances of building elements that needed to be classified to the ontological classes as obtained when formalizing the *Decret*. We implemented the extraction algorithm proposed by Zhang [4]. The main idea is to populate the ontology with instances from an enriched IFC file. For each IFC class the algorithm gets the IFC property set of each instance and adds it into the corresponding ontology class. Instances are identified by a Globally Unique Identifier (GUID) and the property set of each instance is added to the ontology as data or object properties.

2.3. Execution of the rule-checking process

The classification task was done with a DL reasoner (HermiT v.1.3.8 [5]) by distinguishing those instances that are classified as valid according to a rule, from those that are classified as invalid because of some violation of a rule, and from those that are classified as lacking data. Figure 2 shows in light yellow the classification of BIM instances to the *Decret* ontology. These instances comply with Rule 3.15.

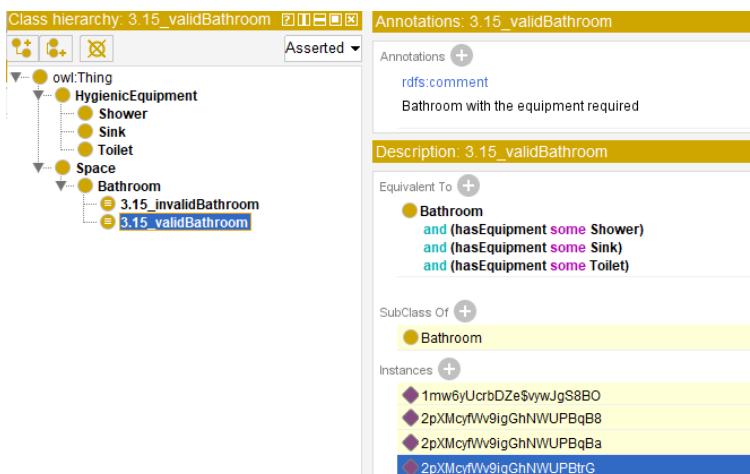


Figure 2. Example of ontology classification and instance evaluation (Rule 3.15).

2.4. Reporting back the checking results

Our approach provides two outputs for the end users, that is, a text and a visual format. The text format contains a data list. Every row of the list describes the instance GUID, the

⁴IFC file owner, <http://www.eipm.es/es/>

number of the evaluated rule, a tag that indicates the validation status, and the compliance description. The visual format is an IFC graphic representation. The visualiser gets the list of the text format, and classifies and shows the 3D instances coloring them according to the compliance status with respect to the regulation, namely, green if valid, red if invalid, and yellow if lacking data. Figure 3 shows the evaluation results for Rule 3.15. a) contains a list of bathrooms and c) the visualization of five valid spaces.



Figure 3. Results: a) text format and b) graphic format.

3. Challenges

During the engineering of the *Decret* ontology for the building regulations on habitability, and the application of DL reasoners in order to check the compliance of particular BIM models with respect to the *Decret*, we encountered three main challenges: Extracting formal definitions from natural language, linking IFC specifications with ontologies, and checking geometric conditions. In this section we describe how we have addressed them:

3.1. Extracting formal definitions from natural language

Extracting formal definitions from natural language text is a challenging task due to the lack of mature natural language processing techniques. In addition, regulations may have several meanings, may be vague, may have ambiguities, and may have references to internal and external norms. Building regulations also have those problems. Let's consider Example 3, which provides additional details about hygienic equipment and bathrooms.

Example 3. Annex 2, Rule 6.4:⁵ Have hygienic equipment, understood as hygienic appliances that, with the corresponding running water and drainage, are intended for hygiene and evacuation of the human body, so that:

- a) It consists of at least one sink, one toilet, and either a shower or bathtub, all in good conditions.
- b) The toilet must be included in a bathroom that can be made independent.

⁵“Disposar d'un equip higiènic, entès com els aparells higiènics que, amb la dotació d'aigua corrent corresponent i el desguàs, estan destinats a la higiene i l'evacuació del cos humà, de manera que:

- a) Estigui format com a mínim per un lavabo, un vâter i una dutxa o banyera en bon estat.
- b) El vâter ha d'estar inclòs en una cambra higiènica independitzable.”

In a), the phrase *in good condition* lacks a clear definition, and so does the phrase *can be made independent* in b). Construction experts describe the latter as being isolated from other spaces by having walls or partitions as well as being accessible, for example, through a door.

In order to avoid ambiguity and facilitate the translation of building regulations into a formal representation such as description logics, we propose to use representations in some controlled natural language (CNL). We have explored Attempto Controlled English (ACE) and its sublanguage for representing description logics [6], but we have realised that current state-of-the-art CNLs are not expressive enough for many of the rules of building regulations such as the *Decret*. For instance, the representation of comparison operators and their numerical data are not supported by ACE, e.g. “height must not be less than 2.20 m” (Annex1, Rule 3.5).

3.2. Linking IFC specifications with ontologies

The second challenge encountered was linking low-level IFC building specifications with high-level ontological concepts obtained from normative text. IFC specifications are based on standard classes, but they do not describe all the detailed semantics of buildings that is needed for automatically checking their compliance with respect to building regulations. For example, Figure 4 shows an extract of an IFC specification that defines an instance with ID=#17185 using the IfcSpace class. IFC does not provide standard classes that define this instance explicitly as a bathroom.

```
#17185= IFCSPACE('2pXMcyfWv9igGhNWUPBtrG',#42,'CH',$,,$,#17171,#17182,
'CH Cambra Higi\X2\00E8\X0\nica', .ELEMENT..,INTERNAL.,$);
...
#17357= IFCPROPERTYSINGLEVALUE('ACO_Class_GuBIMclass',$/,
IFCTEXT('[Uniclass 2015 SL 1.19] SL_35_80_08: Bathrooms'),$);
#17358= IFCPROPERTYSET('3EUBrWARjF5vYGum69v2OM',#42,'INCASOL_IDN','','(#17357)');
...
#17385= IFCRELDEFINESBYPROPERTIES ('07EPENH6bDi9T6Qfsv7eaC',#42,$,$,(#17185),#17358);
```

Figure 4. Extract of a real IFC file.

We propose to enrich the BIM model by adding properties and classification systems by inserting explicit information about IFC building elements. Figure 4 shows in the entity with ID=#17357 a property-value (the classification), which is linked to the instance with ID=#17185 in the IfcRelDefinesByProperties (ID=#17385). In particular, the entity with ID=#17357 holds a Uniclass⁶ code SL_35_80_08 that states that we are dealing with a bathroom. In the extraction process of IFC data, all the explicit bathroom spaces will be classified under the Bathroom class of our ontology. Figure 5 shows the ID=#17185 instance as of class Bathroom, and its data properties in green (e.g., height, etc.) and its object properties in blue (three hygienic appliances).

⁶Uniclass is a unified classification system for the construction industry (<https://uniclass.thenbs.com>).

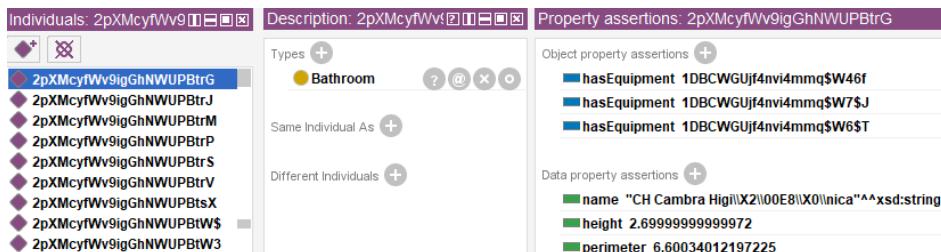


Figure 5. Example of a bathroom instance.

3.3. Checking geometric conditions

Some rules of building regulations require to evaluate geometrical and spatial relations between building entities. In the *Decret*, we found seven of such rules. Example 4 shows one of such type of rule, namely the one that defines when a space is practicable, which means that people with reduced mobility can access the relevant elements of the space (e.g., the bed in a bedroom).

Example 4. Annex 1, Rule 3.4.1.e.⁷ *In practicable spaces, it should be possible to inscribe a circle of one meter and twenty centimeters (1.20 m) in diameter, free of the impact of the rotation of doors and of fixed equipment up to 0.70 m high (toilets and furniture). The interior routes of these spaces must have a minimum passage width of 0.80 m.*

When reading this rule, humans deduce that the interior routes of a space are required in order to allow access to the main elements of this space (e.g., a bed or wardrobe in a bedroom). Construction experts require that the access side of each element is specified, for example, the wardrobe's door would be its access side.

The compliance check of a geometrical norm with respect to an ontology is possible using external geometry algorithms and spatial reasoners. In other words, by means of geometric analysis and spatial processing on the IFC file, we update concrete ontological values of our populated ontology so as to reflect the compliance or not of the geometrical norm. Therefore, to provide a solution for addressing the geometrical requirements of the rule of Example 4, we propose employing geometrical algorithms from the CGAL library⁸ for spaces in two dimensions, 2D polygons automatically extracted from IFC file. To check the requirement of the inscription of a circle, we use the skeleton algorithm for 2D polygons. The idea is to obtain a set of internal vertices, and to calculate circles of 1.20 m, with its secants and tangents as produced by the edges of the polygons. Our check looks of the absence of secants and lines inside the circle (i.e., it prevents a wall from crossing the circle). If the algorithm finds a valid circle then the ontology property of the instance is updated to reflect the compliance with the geometric condition. Figure 4 a) illustrates a 2D view of a bedroom, where the fixed equipment is in green and the inscribed circle in red.

⁷“En els espais practicables s’ha de poder inscriure un cercle d’un metre i vint centímetres de diàmetre (1,20 m), lliure de l’afectació del gir de les portes i dels equipaments fixos de fins a 0,70 m d’alçada (sanitaris i mobiliari). Els recorreguts interiors d’aquests espais han de tenir una amplada mínima de pas de 0,80 m.”

⁸<https://www.cgal.org>.

A similar solution is implemented for the internal route. First, we employ the offset algorithm that generates a new internal area from an initial 2D polygon that shows the areas that are wider than the required passage width. We can then verify if the route has access to the fixed equipment delimited by its access sides (determined by construction experts in the design). The main door creates an access side using a bisector strategy considering the door segment. If the route connects all the access sides, then the ontology property is updated to reflect the compliance with the geometric condition. Figure 6 b) shows a 2D view of a bedroom, where the equipment is in green, an internal route is in blue, the main door segment is in orange, and access sides are indicated by the red arrow lines.

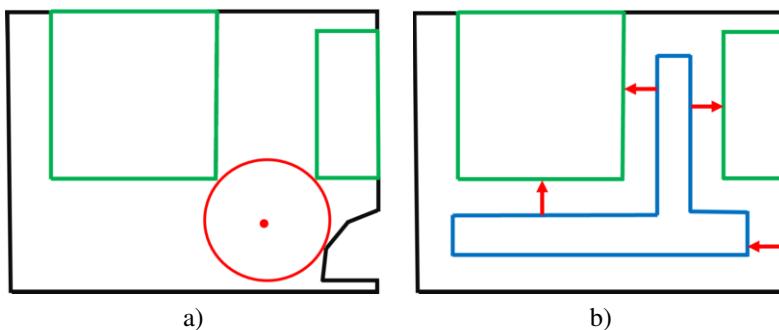


Figure 6. Geometric requirements: a) inscription of a circle b) internal route.

4. Related work

Hjelseth and Nisbet suggested a methodology for marking up the expressions of a rule according to four categories: Requirements and Applicability (i.e., what requirements apply to which entities), and Select and Exceptions (i.e., alternative subjects or other exceptions to the rule) [7]. They proposed a way to re-formulate norms based on this mark-up, and developed a software to check for rule compliance, where the 2021 version uses a conceptual graph to represent the semantics of rules. Our approach represents rules as definitions in an OWL2 ontology, and we explored the use of controlled natural languages for rule specification.

Zahng and El-Gohary use a rule-based, semantic natural language processing approach to extract norms in first-order logic from various construction regulatory documents in order to develop a compliance checking system [8,9]. In contrast, we do not focus on natural language processing, given the ambiguity of regulations, and advocate for rules to be written in some controlled natural language first, prior to extraction and formalization. In the classification of ontology instances, we use then an enriched BIM model.

A cloud-based solution to check rules called BIM-kit is proposed in [10]. The authors describe two interesting services, the rule editor and the model checking service. The first service allows the user to write the rules in a graphical mode, a restricted natural language, or a code representation system. The second service checks the regulation by

employing an external app that returns the resulting list to a 3D visualizer. In our work, we use the ACE plugin⁹ to write the regulation, with which we edit, view and create OWL2 ontologies.

Doukari et al. also proposed a framework for automated compliance checking based on BIM [11]. In the rule interpretation, they mention the immaturity of natural language processing technologies, and as a consequence, they code the regulation into if-then logical rules in C#. The BIM model is loaded as a data structure of partial states. Next, the rule and the building element are concatenated and evaluated as ‘pass’ or ‘fail’. The last step is a detailed report. Our approach offers a similar process with some variations, for example, using a controlled natural language that generates definitions (classes) of an ontology. From the BIM model we extract the enriched instances to populate and evaluate an OWL2 ontology, so as to make it reusable on the web. And, in addition to identifying those building elements that ‘pass’ or ‘fail’, we also categorize those that lack data to make a definite decision on their validity.

There are methods to evaluate geometrical errors (of design and modeling) and regulations concerning spatial aspects. Dinis et al. proposed a Virtual Reality check tool with which the end-users have an immersive experience in a 3D representation to find errors and to subsequently explain their findings [12]. From our experience, this evaluation is time consuming and requires experienced users. In order to evaluate spatial data of a BIM model, it would be possible to use a formal query language such as BimSPARQL [13]. In our work, we evaluate the spatial aspects of norms by resorting to libraries of computational geometry algorithms such as CGAL.

Recently, neuronal network approaches have also been explored to evaluate housing constructions based on a set of norms [14]. The authors formalize the regulation specified manually by architects to create the network. In contrast, we chose to follow a logic-based approach that does not need a training stage, as with learning techniques. In turn, we require the regulators to be familiar with controlled languages so as to be able to write the norms in a more constrained and unambiguous way, or to work collaboratively with specialists in the translation from the regulators’ natural language to the controlled one.

Finally, a popular commercial software is the Solibri Model Checker.¹⁰ Version 2022 has 56 single rule templates to evaluate IFC models. These templates define standard checking procedures, delimited by a number of parameters. The end-users can edit the rule templates according to their requirements. To formulate the rule templates, there is a need for experienced users with an extensive knowledge of IFC.

5. Conclusions and future work

In this paper, we presented a prototype a case study for automated compliance checking of building regulations. First, we formalized a fragment of Catalonia’s *Decret d’habilitat*, which regulates the habitability requirements for buildings; we did this supported by a controlled natural language, in order to define an OWL2 ontology that captured the requirements for building elements as specified in the *Decret*. Next, we enriched a BIM model as described using IFC employing standards of classification systems such as Uniclass. This meant adding explicit information to the various building elements specified

⁹<http://attempto.ifi.uzh.ch/aceview/>

¹⁰<https://www.solibri.com>

in the BIM model. Then, we implemented an extraction algorithm for the enriched BIM model to populate the *Decret* ontology, and used a DL reasoner to classify the instances of the model to the ontology classes. In the last step, we generated reports for end-users, either in text or in a graphical mode, showing those instances of building elements that are correctly classified or are incorrectly classified because either they violate one or more rules of the regulation, or else lack data. As a future work, we would like to set up an empirical evaluation of our proposal using the entire *Decret*, checking the compliance of BIM models of different IFC file size, in order to evaluate our proposal in terms of expressiveness and computational cost. We also plan to study natural language processing techniques that are suitable for formalizing building regulations, seen as mathematical word problems [15].

References

- [1] C. Eastman, P. Teicholz, R. Sacks, K. Liston, *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors*, John Wiley & Sons, 2008.
- [2] Generalitat de Catalunya, Portal Jurídic de Catalunya, DECRET 1412012, de 30 d'octubre, pel qual es regulen les condicions mínimes d'habitabilitat dels habitatges i la cedula d'habitabilitat, <https://portaljuridic.gencat.cat/eli/es-ct/d/2012/10/30/141>, accessed:2021-01-15 (2012).
- [3] C. Eastman, J. Lee, Y. Jeong, J. Lee, Automatic rule-based checking of building designs, *Automation in Construction* 18 (8) (2009) 1011–1033.
- [4] L. Zhang, N. M. El-Gohary, Extracting Information from Building Information Models to Support Automated Value Analysis (2016) 527–534.
- [5] R. Shearer, B. Motik, I. Horrocks, Hermit: A highly-efficient OWL reasoner, in: Proceedings of the 5th OWL: Experiences and Directions Workshop (OWLED 2008), 2008.
- [6] K. Kaljur, N. E. Fuchs, Verbalizing OWL in Attempto Controlled English, in: In Proceedings of OWLED07, 2007.
- [7] E. Hjelseth, N. N. Nisbet, Capturing normative constraints by use of the semantic mark-up rase methodology, 2011, pp. 1–10.
- [8] J. Zhang, N. M. El-Gohary, Automated information transformation for automated regulatory compliance checking in construction, *Journal of Computing in Civil Engineering* 29 (2015) 1–16.
- [9] J. Zhang, N. M. El-Gohary, Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking, *Automation in Construction* 73 (2017) 45–57.
- [10] C. Sydora, E. Stroulia, BIM-kit: An extendible toolkit for reasoning about building information models, in: Proceedings of the 2021 European Conference on Computing in Construction, Vol. 2 of Computing in Construction, University College Dublin, Online, 2021, pp. 107–114.
- [11] O. Doukari, D. Greenwood, K. Rogage, M. Kassem, Object-oriented compliance checking: an automated approach and a case study, in: Proceedings of the 2021 European Conference on Computing in Construction, Vol. 2 of Computing in Construction, University College Dublin, Online, 2021, pp. 293–302.
- [12] F. M. Dinis, J. Martins, F. Sousa, B. Rangel, A. Guimarães, A. Soeiro, Virtual reality design quality-check tool for engineering projects, in: Proceedings of the 38th International Conference of CIB W78, 2021, pp. 912–922.
- [13] C. Zhang, J. Beetz, de Vries, BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data, *Semantic Web* (2018) 1–27.
- [14] H. Arora, J. Bielski, V. Eisenstadt, C. Langenhan, C. Ziegler, K.-D. Althoff, A. Dengel, Consistency checker an automatic constraint-based evaluator for housing spatial configurations, in: International conference eCAADe – Education and research in Computer Aided Architectural Design in Europe, 2021, pp. 351–358.
- [15] S. Ughade, S. Kumbhar, Survey on mathematical word problem solving using natural language processing, in: 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1–5.

Deep Neural Classification of Darknet Traffic

Mahmoud Alimoradi^a, Mahdieh Zabihimayvan^{b,1}, Arman Daliri^c, Ryan Sledzik^d, and Reza Sadeghi^e

^{a, c} Independent researchers

^{b,d} Department of Computer Science, Central Connecticut State University, New Britain, CT, USA

^e School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY, USA

Abstract. Darknet is an encrypted portion of the internet for users who intend to hide their identity. Darknet's anonymous nature makes it an effective tool for illegal online activities such as drug trafficking, terrorist activities, and dark marketplaces. Darknet traffic recognition is essential in monitoring and detection of malicious online activities. However, due to the anonymizing strategies used for the darknet to conceal users' identity, traffic recognition is practically challenging. The state-of-the-art recognition systems are empowered by artificial intelligence techniques to segregate the Darknet traffic data. Since they rely on processed features and balancing techniques, these systems suffer from low performance, inability to discover hidden relations in data, and high computational complexity. In this paper, we propose a novel decision support system named Tor-VPN detector to classify raw darknet traffic into four classes of Tor, non-Tor, VPN, and non-VPN. The detector discovers complex non-linear relations from raw darknet traffic by our deep neural network architecture with 79 input artificial neurons and 6 hidden layers. To evaluate the performance of the proposed method, analyses are conducted on a benchmark dataset of DIDarknet. Our model outperforms the state-of-the-art neural network for darknet traffic classification with an accuracy of 96%. These results demonstrate the power of our model in handling darknet traffic without using any preprocessing techniques, like feature extraction or balancing techniques.

Keywords. Darknet traffic, Machine learning, Decision support system, Deep neural network, Tor, Classification.

1. Introduction

Anonymity networks complicate any possibility of tracking and tracing of users' identity on the Web and rely on a worldwide network of volunteer Web servers. Darknets such as Tor and I2P are anonymity networks that prevent traffic analysis and activity monitoring using encryption schemes like onion routing [1]. The anonymity on darknets is indeed provided for both senders and receivers. This anonymous nature allows users to carry on illegal activities as dark hidden services. A web of such services on darknets such as Tor is called dark Web and there has been a great deal of work to analyze the content and application of hidden services on dark Web [2] [3]. However, the focus of this paper is on classification of network traffic on darknets, rather than investigation of dark Web.

¹ Corresponding Author; E-mail: zabihimayvan@ccsu.edu.

Darknet traffic classification plays an important role in detection of cyberattacks and malicious activities on the Internet [4] [5]. There have been significant efforts to detect and classify encrypted traffic of different darknets that rely on machine learning techniques. Hu et. al. propose a hierarchical classification method to identify the type of traffic (darknet or regular internet), type of darknet (Tor, I2P, ZeroNet, Freenet), and user behavior on each network [6]. Choorod and Weir propose a character frequency approach to classify Tor traffic based on characteristics of the encrypted payload. They employ and evaluate different machine learning methods to distinguish Tor packets from regular Web traffic [7]. However, there is few studies on evaluation of deep neural networks to detect and characterize darknet traffic. Perhaps our best understanding of deep neural networks as darknet classifiers is from Lashkari et. al. [8], who proposed a method based on convolutional neural networks to classify darknet traffic. Their method utilizes a feature selection technique to find the most important features and create a gray image that is fed into a two-dimensional convolutional neural network to detect and characterize traffic. Following motivations led to our study:

- The data that is used for darknet traffic classification should be a recent benchmark data that not only can be accessible by other researchers, but also contain both anonymized VPN and Tor activity traffic to represent the real darknet traffic.
- Darknet traffic data contains a large list of features for traffic samples and many related studies employ different feature selection techniques to reduce the number of features to a number that is manageable by existing machine learning methods.
- Highly imbalanced data is naturally inherent in cybersecurity applications such as darknet traffic classification, fraud, and phishing attack detection [9]. This can pose difficulty and inefficiency to machine learning methods due to bias towards majority class. However, balancing data using oversampling (or undersampling) can discard useful information about the data that can be crucial for classification [10].

In this paper, we propose a novel deep neural network to classify traffic data into four classes of Tor, VPN, non-Tor, and non-VPN. The experiments rely on a large dataset that is recently collected and published on Kaggle [8]. We propose a deep neural network as the classifier to distinguish between regular and darknet traffic. Our model can also handle the high-class imbalance without any preprocessing technique to balance the data. The experimental results indicate that our proposed deep neural network outperforms the state-of-the-art deep neural network for darknet traffic classification with the accuracy and F1 of 96%, and Kappa value of 0.92. The neural network we propose in this work can also identify salient features in traffic data with no need for a feature extraction technique prior to detection.

The rest of this paper is organized as follows: in Section 2, we first discuss the related work on darknet traffic characterization and classification. Section 3 provides a background knowledge on deep neural networks and describes the network proposed in this study. Section 4 presents the experiments to evaluate the performance of the proposed model and Section 5 discusses the conclusion and future direction of our work.

2. Related Work

There has been recently a great deal of effort on darknet traffic classification although the emergence of Web traffic classification backs to two decades ago [11-15]. Nishikaze et al. propose a new system based on machine learning techniques to monitor malicious activities on the Internet [16]. The packets studied in their work were captured in a communication from a source network to a dark net. For each packet, 27 categories of traffic analysis profile were created in the form of a 27-dimensional feature vector. They used hierarchical clustering to identify malicious packets and matched malware signatures with identified packets. Ban et al. provide a study on early detection of attacks on darknet. They utilize a time series to characterize the activity level of attack patterns [17]. They also reveal the most prominent attack patterns by employing a clustering algorithm that clusters the attack patterns into groups with the same activities. To provide visual insights into the relationships between clusters, a dimension reduction is employed. Their experimental results indicate the effectiveness and efficiency of the proposed approach in early detection of new attack patterns.

To gain a better understanding of the darknet traffic and its parameters in attack identification, Gadhia et al. focused on a comparative analysis over two darknet sensors [18]. Studying total incoming packet, number of source host, targeting destination port for TCP and UDP protocols, they discovered that the darknet sensors have wide difference in incoming traffic characteristics. Fachkha et al. proposed an approach to infer and characterize DNS Distributed Reflection Denial of Service (DRDoS) attacks in dark network [19]. Their work relied on intensity, rate, and geo-location in addition to various network-layer and flow-based features. They employed k-means clustering to identify campaigns of DRDoS Attacks. In another attempt, Fachkha and Debbabi presented a comprehensive survey on darknet and discuss on other trap-based monitoring systems and compare them to darknet [20]. They report case studies on Conficker worm, Sality SIP scan botnet, and the largest amplification attack in 2014 to provide analysis on darknet information. Their work further identifies Honeyd as probably the most practical tool to implement darknet sensors.

Ling et al. proposed and implemented a system to discover and study malicious traffic over Tor [21]. The system uses an intrusion detection system to classify the malicious traffic. Their experimental result reveal approximately 10% of Tor traffic that can trigger alerts of the intrusion detection system. The identified malicious traffic includes P2P traffic, malware traffic, denial-of-service attack traffic, spam, etc. Wang et al. proposed a new person attribute extraction method with the aim of obtaining a comprehensive characterization of malicious users and tracing them [22]. Their method is comprised of block filtration, attribute candidate generation, and attribute candidate verification. Using the extracted information, they analyze sensitive personal information such as Top-K name entities, email domain name, etc. of darknet users.

In [23], three different super-resolution algorithms are used for text recognition in the Darknet. They evaluated their proposed algorithm over five state-of-the-art datasets for text spotting in Tor darknet. Their model achieves a 3.41% of improvement when deep CNN and the rectification network are combined. In [24], the authors utilize unsupervised and self-supervised machine learning methods to infer image semantics from unstructured multimedia data to investigate the content of the Darknet's marketplaces. The evaluation demonstrates how the combination of CNN and LDA models can retrieve documents and images from text and image queries on the Darknet.

In [25], network flow features are used for Tor traffic analysis and multi-level cataloging. Their proposed model can detect the anonymous traffic in different levels of L_1 , L_2 , and L_3 for various platforms such as mobile and PC. The work in [26] and [27] investigates the automatic classification of images on Tor darknet websites. The authors propose a semantic attention keypoint filtering (SAKF) to remove non-significant features at the pixel level of images using a bag of visual words (BoVW) framework to improve the classification accuracy.

3. The Proposed Deep Neural Network

A deep neural network is defined as a tuple $N = (L, C, F)$. $L = \{L_i | 1 < i < M\}$ is a set of M layers where the first layer (L_1) is called input, the last layer (L_M) is called output, and the rest are called hidden layers. C in the tuple is defined as $C = L \times L$ which represents a set of connections between all layers, and $F = \{F_i | 2 < i < M\}$ is a set of functions each of which is used for a non-input layer. Each layer L_i consists of P_{L_i} perceptrons where i^{th} perceptron on layer 1 is denoted by $p_{i,l}$. For each perceptron $p_{i,l}$ in layer L_i , $1 < i < M$, there are two variables $b_{i,l}$ and $a_{i,l}$ that store the values of the perceptron before and after applying an activation function. Activation function or transfer function decides how the value of a perceptron influences its output [28]. The activation function used for hidden layers in this work is ReLU that is the most well-known activation function for deep neural networks. ReLU changes the perceptron's value based on the Equation 1.

$$a_{i,l} = \text{ReLU}(b_{i,l}) = \begin{cases} b_{i,l} & \text{if } b_{i,l} \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

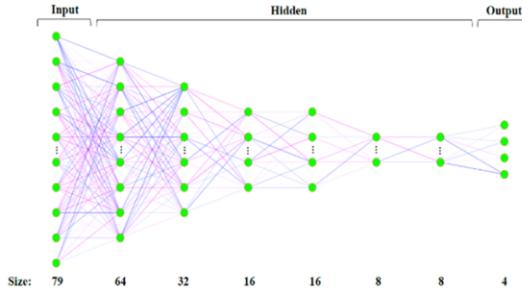
Since there is no activation function to be applied on input (L_1) layer, the perceptrons on the first layer are associated with only one value that is $a_{i,l}$. Each layer L_i in the network is associated with a vector space, $V_{L_i} = \mathcal{R}^{P_{L_i}}$, to record the $a_{i,l}$ values of its perceptrons, and V_{L_1} is considered as an input. In fully connected neural networks such as the network used in this work, all perceptrons in layer L_i are connected to the perceptrons in layer L_{i-1} . The connection between perceptron $p_{i,l}$, i^{th} perceptron on layer 1, and perceptron $p_{j,l+1}$, j^{th} perceptron on layer $l+1$, is denoted as $w_{l,i,j}$. Now, the value of a perceptron before activation function is defined as Equation 2.

$$b_{l+1,i} = \beta_{l+1,i} + \sum_{1 \leq j \leq P_l} w_{l,i,j} \cdot a_{l,j} \quad (2)$$

where $\beta_{l+1,i}$ is called bias for i^{th} perceptron on layer $l+1$. Bias values are tuned during the training, and they help shift the activation function. In Figure 1, we visualize such a network that we proposed to classify darknet traffic. Number of perceptrons in the first layer is equal to the number of non-target features. Number of perceptrons in hidden layers is set based on several experiments on the performance of neural network. Output layer contains four perceptrons for four different classes of Tor, non-Tor, VPN, and non-VPN. Other parameters of the network are listed in Table 1.

Table 1. List of parameters and their values in our deep neural network

Parameter name	Parameter value
Activation Function (Hidden layers)	ReLU
Activation Function (Output layer)	Softmax
Loss Function	Sparse Categorical Cross Entropy
Optimizer	Adam
Epochs	100
Batch Size	64
Validation split	33%

**Figure 1.** The proposed deep neural network

- Activation:** As mentioned earlier, the activation function decides how the weighted sum of input to a perceptron forms its output and eventually the network's output. In this work, we apply ReLU activation function for all hidden layers. To obtain a distribution over the 4 classes in darknet traffic classification, Softmax activation function is used for the output layer. Equation (3) indicates how the Softmax function works for the input vector a [29].

$$\text{softmax}(a) = \frac{e^{a_i}}{\sum_{1 \leq j \leq K} e^{a_j}} \quad (3)$$

where K is the number of classes (4 in our problem), e^{a_i} is the standard exponential function for the input vector, and e^{a_j} is the standard exponential function for the output vector.

- Loss:** Deep neural networks are trained based on stochastic gradient descent or its variants. Loss is the prediction error of the network while calculating and updating the weights. The loss function, sparse categorical cross entropy in this work, is used to calculate loss value of the prediction. Equation 4 indicates how the sparse categorical cross entropy is defined.

$$L(w) = \frac{1}{N} \sum_{k=1}^N [O_k \log(\widehat{O}_k) + (1 - O_k) \log(1 - \widehat{O}_k)] \quad (4)$$

In the equation 4, w indicates the weight vector of the neural network, O_k indicates the true labels of the data, and \widehat{O}_k represents the labels predicted by the network. N also indicates the input size.

- One hot encoder:** This function is used to transform all the categorical data into numerical form which can help the network to have a better prediction [8].

4. **Dropout:** This function is used to prevent overfitting of the network. A dropout layer randomly sets the value of input perceptrons to zero with a frequency of rate (τ) at each step during training the network. Non-zero inputs are scaled up by $\frac{1}{1-\tau}$ such that the sum of all input perceptrons remains the same. In the experiments, we use $\tau = 0.2$ which produces the best results for this problem.
5. **Optimizer:** In this work, we utilize Adam gradient-based optimization algorithm to update the weights of the network during the training phase [30].
6. **Normalization:** To avoid bias the model towards features with high values, normalization is used to transform the features' values into a decimal in the range of $[0, 1]$. In this work, we utilize min-max normalization function, N , that works based on Equation 5.

$$\forall f \text{ in } F: N(f) = \frac{f - f_{min}}{f_{max} - f_{min}} \quad (5)$$

- where F is the set of all features in the data, f indicates a feature, and f_{min} and f_{max} are the minimum and maximum values of f .
7. **Adaptive learning rate:** Each time the network weights are updated during training, learning rate hyperparameter is used to control size of moving towards a minimum of the loss function. An initially large learning rate value helps the model accelerate training and gradually reducing the learning rate helps the model learn complex patterns in the data. In this study, we use exponential decay function [31] that is shown in Equation 6.

$$lr = lr_{in} \times kt \quad (6)$$

In the equation above, lr_{in} is the initial learning rate value, k is a hyperparameter to control the reduction amount, and t is the iteration number. To set the initial learning rate, we consider changes of learning rate in response to the loss value during the training phase.

4. Experiments

We first discuss on the dataset used for darknet traffic classification. This data is called DIDarknet and is a recently collected benchmark dataset of 141,529 traffic instances with 79 features² where 55 features have integer values, and the rest are decimal. The dataset contains four classes of Tor (1,392 samples), VPN (22,919 samples), non-Tor (93,355 samples), and non-VPN (23,863 samples) where class 1 (Non-Tor) and class 4 (VPN) are the majority and the minority, respectively.

To evaluate our model, we randomly allocate 20% of the data for test and 80% for training and validation. The training phase is used to train the model and initialize the weights of the neural network. As Table 1 indicates, 33% of the train data is allocated for validation. The validation set is used after training the model to tune the hyperparameters with the aim of improving the model's accuracy. Figure 2 shows values of loss and accuracy for both train and validation sets.

² For the description of all features, please refer to the original website of the dataset: <https://www.unb.ca/cic/datasets/darknet2020.html>

The loss metric is calculated for both training and validation and shows the sum of errors made on each instance in the training and validation sets. In other words, loss measures how the distribution of a class labels are different from the distribution of labels predicted for the class's instances. During the training phase, loss is used to tune the weights of the neural network and each iteration of the optimization, the loss values imply how well the model behaves. In this work, we employ sparse categorical cross entropy as the loss function.

The accuracy metric is used to measure the model's performance after setting the parameters. It is the measure of how accurate the model predicts the true data. Plot of loss values in Figure 2 demonstrates that after training the model, loss decreases on the validation set which implies the neural network performs better after tuning the model's parameters. According to the plot of accuracy, the neural network learns to predict with an accuracy over 95%, and the prediction accuracy on the validation set is almost the same.

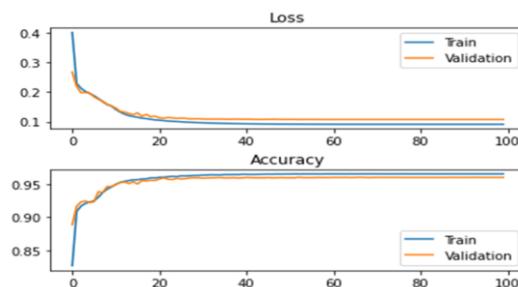


Figure 2. Loss and accuracy for training and validation

As we discussed before, we control the size of the movements on the search space by an adaptive learning rate. To empower our model to use an effective learning rate, we choose its initial value by examining the effects of various learning rates on the loss function. Figure 3 shows the plot of changes where for learning rate values greater than 10^{-3} and close to 10^{-2} , loss is minimum. By guess and check iterations over the values in this range, we set the initial learning rate to 3×10^{-3} .

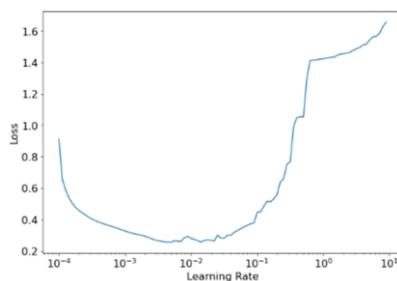


Figure 3. Loss Vs. Learning rate

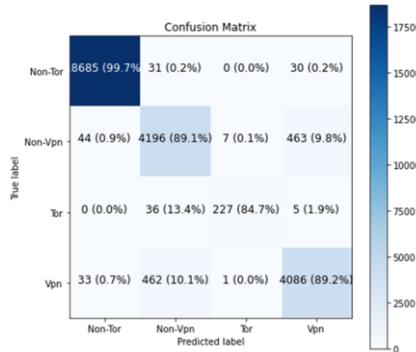
We now evaluate the deep neural network's performance using the test data. The data is fed into the input layer, and the evaluation metrics listed in Table 2 are calculated. Since the data in this work contains more than two classes, all the evaluation metrics in Table 2 are used as multi-class metrics. Also, for an easier interpretation, values for the first four metrics are reported as percentage.

Table 2. Evaluation metrics used to evaluate the deep neural network's performance

Metric	Equation	Metric variables	Value
Accuracy	$\frac{TP + TN}{P + N}$	P : size of class P N : size of class N	96.07
Precision	$\frac{TP}{TP + FP}$	TP : No. of samples correctly classifies as P FN : No. of samples incorrectly classifies as P	96.08
Recall	$\frac{TP}{TP + FN}$	TN : No. of samples correctly classifies as N FP : No. of samples incorrectly classifies as N	96.12
F1	$\frac{2TP}{2TP + FP + FN}$	FN : No. of samples incorrectly classifies as N	96.06
Kappa	$\frac{P_{obs} - P_{exp}}{1 - P_{exp}}$	P_{obs} : empirical probability of agreement on the label assigned to any sample P_{exp} : expected agreement when labels are assigned randomly	0.9225

Accuracy is one of the well-known metrics used to evaluation classification performance of machine learning techniques. Accuracy reports total number of correct predictions to the total number of all samples. In case of imbalanced data, accuracy is not a proper metric since the ratio can be biased towards the majority class. Precision is another classification metric that indicates how precise the classifier is in predicting true samples of each class. In other words, precision reports what portion of all samples classified in a class truly belong to that class. In contrast, recall is a metric to represent the percentage of samples in a class that are correctly predicted by the classifier. F1 is the harmonic mean of precision and recall and indicates the quality of classification.

In contrast to the stet-of-the-art model, DIDarknet, [8] with accuracy of 85%, our model notably outperforms based on four well-known evaluation metrics. To gain a better understanding of the model performance, the confusion matrix of the model is shown in Figure 4. The diagonal values indicate how well the model predicts true data in each class while non-diagonal values on each row indicate number of instances in a class that are incorrectly classified in other classes.

**Figure 4.** Confusion matrix of the classification

Kappa is a statistical score in $[0,1]$ that reports how much two annotators (true labels and predicted labels) agree on the labels assigned to the samples in a classification problem. The following guideline published in [32] can be used to interpret the Kappa metric: value 0.00 to 0.20 is considered slight agreement; 0.21 to 0.40 is fair agreement; 0.41 to 0.60 is moderate agreement; 0.61 to 0.80 is substantial agreement; and 0.81 to 1.00 is almost perfect agreement. According to the value reported for Kappa in Table 2,

the deep neural network presents a perfect agreement between the true and predicted labels in darknet traffic classification.

5. Conclusion and Future Work

Darknet traffic classification plays an important role in detection of cyberattacks and malicious activities on the Internet. This work proposes a deep neural network for darknet traffic classification. We utilize a recently published benchmarked dataset of Web traffic that contains both anonymized VPN and Tor activity instances to represent the real darknet traffic. The main purpose is to classify the darknet traffic into four classes of non-Tor, non-VPN, Tor, and VPN. The state-of-the-art deep neural models proposed for the problem employ feature selection/extraction prior to classification to reduce number of features. However, our model can identify salient features in traffic data during training the network. It also handles the high-class imbalance in the data without any balancing technique prior to classification. Based on different types of evaluation metrics, our model outperforms the related work with a notable difference of 10% in classification accuracy and Kappa value of 0.92.

To extend this work, we plan to evaluate the performance of deep neural network in classification of dark hidden services regarding their textual information and their structural identity [33]. Also, creating and publishing another dataset that contains more balanced data of Tor and VPN traffic can be another direction for future work. Regarding rapidly changing traffic of darknet, we can also expand this work further by studying the evolution of the traffic and its features over time and implement a deep neural network to classify big data of darknet traffic over time.

References

- [1] D. Goldschlag, M. Reed and P. Syverson, "Onion routing for anonymous and private internet connections," *Communications of the ACM*, vol. 42, no. 2, p. 5, 1999.
- [2] M. W. Al-Nabki, F. Eduardo and A. Enrique, "Torank: Identifying the most influential suspicious domains in the tor network," *Expert Systems with Applications*, vol. 123, pp. 212-226, 2019.
- [3] M. Zabihimayvan, R. Sadeghi, D. Doran and M. Allahyari, "A broad evaluation of the Tor English content ecosystem," in *Proceedings of the 10th ACM Conference on Web Science*, 2019.
- [4] N. Hashimoto, S. Ozawa, T. Ban, J. Nakazato and J. Shimamura, "A Darknet Traffic Analysis for IoT Malwares Using Association Rule," in *Conference on Big Data and Deep Learning*, Procedia Computer Science, 2018.
- [5] K. Kanemura, K. Toyoda and T. Ohtsuki, "Identification of darknet markets' bitcoin addresses by voting per-address classification results," in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2019.
- [6] Y. Hu, F. Zou, L. Li and P. Yi, "Traffic Classification of User Behaviors in Tor, I2P, ZeroNet, Freenet," in *19th International Conference on Trust, Security and Privacy in Computing and Communications*, 2020.
- [7] P. Choorod and G. Weir, "Tor Traffic Classification Based on Encrypted Payload Characteristics," in *National Computing Colleges Conference*, 2021.
- [8] A. Habibi Lashkari, G. Kaur and A. Rahali, "DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning," in *The 10th International Conference on Communication and Network Security*, 2020.
- [9] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1-54, 2019.
- [10] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42-47, 2012.
- [11] J. P. Early, C. E. Brodley and C. Rosenberg, "Behavioral authentication of server flows," in *Annual Computer Security Applications Conference*, 2003.
- [12] W. H. Turkett Jr, A. V. Karode and E. W. Fulp, "In-the-Dark Network Traffic Classification Using Support Vector Machines," in *Association for the Advancement of Artificial Intelligence*, 2008.

- [13] L. Bernaille, R. Teixeira and K. Salamatian, "Early application identification," in Proceedings of the ACM CoNEXT conference, 2006.
- [14] L. Bernaille and R. Teixeira, "Early recognition of encrypted applications," in International Conference on Passive and Active Network Measurement, 2007.
- [15] C. V. Wright, F. Monroe and G. M. Masson, "On inferring application protocol behaviors in encrypted network traffic," Journal of Machine Learning Research, vol. 7, no. 12, 2006.
- [16] H. Nishikaze, S. Ozawa, J. Kitazono, T. Ban, J. Nakazato and J. Shimamura, "Large-scale monitoring for cyber attacks by using cluster information on darknet traffic features," Procedia Computer Science, vol. 53, pp. 175-182, 2015.
- [17] T. Ban, S. Pang, M. Eto, D. Inoue, K. Nakao and R. Huang, "Towards early detection of novel attack patterns through the lens of a large-scale darknet," in Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 2016.
- [18] F. Gadchia, J. Choi and B. Cho, "Comparative analysis of darknet traffic characteristics between darknet sensors," in International Conference on Advanced Communication Technology, 2015.
- [19] C. Fachkha, E. Bou-Harb and M. Debbabi, "Inferring distributed reflection denial of service attacks from darknet," Computer Communications, vol. 62, pp. 59-71, 2015.
- [20] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," IEEE Communications Surveys and Tutorials, vol. 18, no. 2, pp. 1197-1227, 2015.
- [21] Z. Ling, J. Luo, K. Wu, W. Yu and X. Fu, "TorWard: Discovery, Blocking, and Traceback of Malicious Traffic Over Tor," IEEE Transactions on Information Forensics and Security, vol. 10, no. 12, pp. 2515-2530, 2015.
- [22] M. Wang, X. Wang, J. Shi, Q. Tan, Y. Gao, M. Chen and X. Jiang, "Who are in the Darknet? Measurement and Analysis of Darknet Person Attributes," in Conference on Data Science in Cyberspace (DSC), 2018.
- [23] P. Blanco-Medina, E. Fidalgo, E. Alegre and F. Jámez-Martino, "Improving Text Recognition in Tor darknet with Rectification and Super-Resolution techniques," in IET Conference Proceedings. The Institution of Engineering & Technology, 2019.
- [24] A. Berman and C. L. Paul, "Making sense of darknet markets: Automatic inference of semantic classifications from unconventional multimedia datasets," in International Conference on Human-Computer Interaction, 2019.
- [25] L. Wang, H. Mei and V. S. Sheng, "Multilevel identification and classification analysis of Tor on mobile and PC platforms," IEEE Transactions on Industrial Informatics, vol. 17, no. 2, pp. 1079-1088, 2020.
- [26] E. F. Fernandez, R. A. V. Carofilis, F. J. Martino and P. B. Medina, "Classifying Suspicious Content in Tor Darknet," arXiv preprint arXiv:2005.10086, 2020.
- [27] E. Fidalgo, E. Alegre, L. Fernández-Robles and V. González-Castro, "Classifying suspicious content in tor darknet through Semantic Attention Keypoint Filtering," Digital Investigation, vol. 30, pp. 12-22, 2019.
- [28] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.
- [29] R. Szeliski, Computer vision: algorithms and applications, Springer Science and Business Media, 2010.
- [30] F. Chollet, Deep Learning with Python, Simon and Schuster, 2017.
- [31] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in Neural networks: Tricks of the trade, Springer, 2012, pp. 437-478.
- [32] R. Landis and K. Gary, "The measurement of observer agreement for categorical data," Biometrics, pp. 159-174, 1977.
- [33] M. Zabihimayvan, R. Sadeghi, D. Kadariya and D. Doran, "Interaction of Structure and Information on Tor," in International Conference on Complex Networks and Their Applications, 2020.

Enabling Reproducibility in Group Recommender Systems

Joaquin Dario SILVEIRA ^a, Maria SALAMÓ ^{b,1}, and Ludovico BORATTO ^c,

^a Universitat Politècnica de Catalunya, Barcelona, Spain

^b Dept. Mathematics and Computer Science, University of Barcelona, Barcelona, Spain

Institute of Complex Systems, University of Barcelona, Barcelona, Spain

^c Dept. Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

Abstract. Reproducibility is a challenging aspect that considerably affects the quality of most scientific papers. To deal with this, many open frameworks allow to build, test, and benchmark recommender systems for single users. Group recommender systems involve additional tasks w.r.t. those for single users, such as the identification of the groups, or their modeling. While this clearly amplifies the possible reproducibility issues, to date, no framework to benchmark group recommender systems exists. In this work, we enable reproducibility in group recommender systems by extending the LibRec library, which stands out as one of the richest, with more than 70 different recommender algorithms, good performance and several evaluation metrics. Specifically, we include several approaches for all the stages of group recommender systems: group formation, group modeling strategies, and evaluation. To validate our framework, we consider a use-case that compares several group building, recommendation, and group modeling approaches.

Keywords. Group Recommender Systems, Reproducibility, Algorithms

1. Introduction

Enabling reproducibility should be of paramount importance inside the research community [1]. In fact, it is hard to determine the speed of progress, or even if we are making any, when so much of the newly generated knowledge is not reproducible [2]. The existence of base libraries with known and well studied approaches and algorithms is one of the first steps in any field that seeks to advance on firm knowledge. Moreover, it is in fields in which such frameworks are missing that it is hardest to justify new ideas by benchmarking them against the existing literature.

Recommender systems (RSs) support users by suggesting items that might be of interest to them [3]. This is usually done by learning behavioral patterns from historical data, usually in the form of user-item interactions. Nearly any popular programming language has a library or framework for making single recommendations. Despite the amount of papers regarding the problem of generating recommendations for groups of users (*group recommender systems*, GRSs) [4,5], a firm ground for GRSs does not exist.

¹Corresponding Author: Department of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain; E-mail: maria.salamo@ub.edu

This is exacerbated by the difficulty of accessing to or generating datasets that gather information of actual group recommender systems [6,7,8,9]. Thus, in the RS research field, a common framework for benchmarking GRSs is a known open issue.

In contrast to single RSSs, in a framework for GRSs several issues appear, mainly because a GRS provides suggestions in contexts in which more than one person is involved in the recommendation process and their aim is to provide recommendations to the whole group, considering the preferences and the characteristics of more than one user. Because of this, a great amount of researchers resort to individual recommendation datasets for offline testing and benchmarking [10,11,12,13]. As a consequence, this introduces an important issue regarding the need to *form groups* to whom propose group recommendations. In addition, it would certainly be necessary to address the issue of how to *evaluate the results for groups*. Every group recommendation study seems to tackle these questions differently. Due to these two issues and other factors, there is a whole myriad of ways in which group recommendations can be performed. Furthermore, the development of new strategies and ongoing research in several of these issues makes the task of encompassing all of them in a single framework daunting. For this reason, enabling reproducibility in group recommender systems is of central importance.

In this paper, we enable reproducibility in group recommender systems by extending the LibRec (i.e., www.librec.net) library, which is one of the most widely used recommendation frameworks. The proposed extension encompasses different interpretations of the aforementioned issues. In particular, we focus on several of the stages of group recommender systems: group formation, group modeling, and evaluation. To narrow down the scope of this paper, we tackle the following aspects: (1) *Building of synthetic groups*, where we focus on offline group recommendation with synthetic group formation, due to the lack of real group recommendation datasets; (2) *Single user prediction aggregation*, where we implement the most common approaches of aggregating user preferences [14]; (3) *Measuring members satisfaction with the group recommendation*, where we compare the individual preferences expressed in the test set with the group recommendations.

Our contributions are summarized as follows: (1) We propose a framework² to enable reproducibility in group recommender systems; (2) We elevate some reproducibility questions often not regarded or ignored, which are relevant for understanding and comparing group recommendation experiments; (3) In the scope of collaborative filtering approaches, we enable the use of several combinations in three main stages (group building, group modeling, and evaluation) of group recommender systems; and (4) We present a use-case that compares group building, recommendation, and group modeling approaches, both in terms of effectiveness and efficiency.

2. Framework for Benchmarking Group Recommender Systems

2.1. Introduction to the group recommendation pipeline

The group recommendation framework, called GroupLibRec, is an extension of the LibRec library. We have chosen LibRec as base library because it is under the GPL license, it has already implemented more than 70 recommendation algorithms, has good performance and is widely used in the recommender system research community. Note

²<https://github.com/panserbjorn/librec/tree/3.0.0/RecSys>

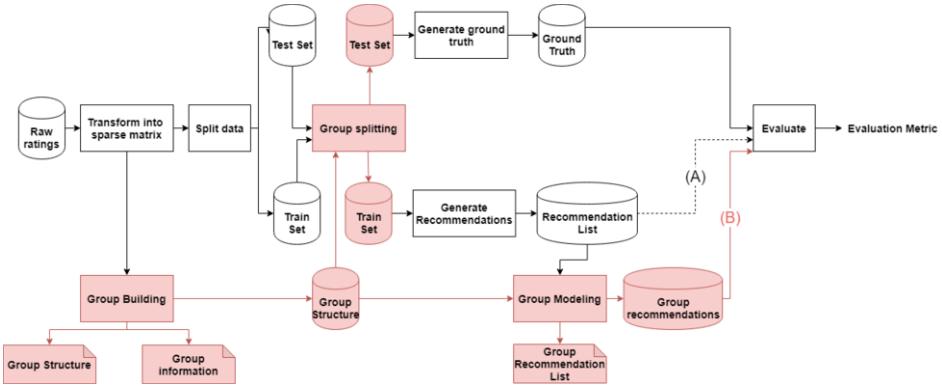


Figure 1. GroupLibRec recommendation pipeline. In red the extensions to the data flow pipeline made to LibRec to enable group recommendations.

that the usage of a base library for building the framework permits us to reuse existing and tested solutions to the data processing and single users recommendation scenarios. This is highly desirable, since most group recommender systems include in their process single user recommendations. The library extension could be pursued in different ways, i.e., as an external project, as another project inside the library or integrating the group recommendations to the core library. We have opted for the last one, mainly because our goal has been to include the group recommendation capabilities with as few changes as possible to LibRec. We consider it will increase the chances of the group capabilities being introduced to stable future versions of the library.

Figure 1 depicts the recommendation pipeline for GroupLibRec, which includes the LibRec pipeline (elements in white color). Note that the LibRec library pipeline performs the steps of data ingestion, splitting it into train and test sets, training a recommender, generating recommendations and, finally, evaluating the results. Configuration files enable the setup of all the required parameters in a run, such as the recommendation algorithm selection, the dataset location, and the evaluation metrics. Thus, this process facilitates the reproducibility of experiments just by sharing and exchanging a single configuration file.

The red components in Figure 1 show the extension of the pipeline for enabling group recommendations. The additional steps are: *group building*, *group modeling*, and *splitting*. Specifically, the Group Recommender delegates the responsibility of generating the individual recommendations in the pipeline to the individual user recommender algorithm and then it builds the recommendations for the groups. Specific additions to the recommendation process for enabling group recommendations will be discussed in the following sections.

2.2. Group Building

Due to the lack in datasets that contain natural groups information, most of the research work on group recommendations starts with the formation of synthetic groups inside datasets of individual users ratings. We consider three aspects associated to synthetic groups: (i) if they are overlapping, (ii) if they use internal or external information, and (iii) if the groups cover the entire user base.

First, in relation to the overlaps, *groups that do not overlap* could be stable or occasional [15], e.g., family, colleagues, or close friends. These groups gather to share a common experience and recommendations about such experiences might be sought. This is the case of research studies such as [14] and [16] that propose a system for generating group recommendations regarding trips, or [13] that proposes group recommendation in the context of movies. In these groups, we assume that the similarity between users is high, either because they are gathered together due to those similarities or because their interactions as a group shape their responses to certain items. For the study of this kind of groups, we implemented two approaches, capable of detecting them in individual user datasets, namely *similar users group identification* and a *k-means clustering* technique. On the other extreme, *groups that have big overlaps* could be non established or random groups such as people gathered in the same physical place for an event like a club [15]. Members in these groups might not share much in common and they do not necessarily present a defined inner structure for deciding between different options [17]. The same users that are present in a group might behave differently in another group and therefore could be considered as a different user altogether [7]. That is why we believe that the study of these groups is the most interesting, not because of the overlapping nature, but mainly because of the diversity of the members inside each group. To generate diverse groups we have included a *random group generation* strategy, described below.

Second, considering the usage of *internal or external information*, most of the state-of-the-art approaches use internal information, such as rating matrix. On the other hand, other group strategies use external information such as demographics or social network information [7] for building groups. Currently, the framework focuses on building groups using internal information and similarities between users. The usage of external information will be addressed in the future work.

Finally, the last aspect we consider for group building is the *coverage of users when groups are formed*. The coverage determines if every user inside the dataset belongs to a group or not. The coverage does not rise when the origin of the data has already groups built in, because the assumption of groups being analyzed excludes single users that do not fit into any group. However, when generating synthetic groups it gains more relevance. This arises the problem of what do do when users do not belong to any group, by either removing or maintaining them in the dataset. Our decision was to maintain them since their ratings may contribute to the collaborative filtering process. Thus, we decided to allow formation of groups for a dataset that did not cover all users. We will return on the coverage aspect when discussing the evaluation and splitting.

Overall, the three strategies (i.e, k-means, similar users, and random) in our framework produce non-overlapping groups, which reduces the number of possible groups, but simplifies the evaluation of the group recommendations. To reproduce experiments, both the assignation of users to groups and group information are stored in files. Next, we detail the group building strategies.

K-means based group builder. This strategy assigns a group to every user based on the K-means algorithm applied over the rating matrix. However, since the rating matrix contains sparse data, the Euclidean distance between the centroid and the users is computed only in the common items. For the non shared items, the maximum distance is used. This maximum distance depends on the rating scale present in the dataset. An interesting aspect of the K-means approach is that it can be used to identify a certain number of target groups, independently of their size. Note that the group size depends on the content of

the rating matrix. In contrast, the two approaches we present in what follows are defined based on the size of the groups. These approaches are more useful to study scenarios that are meant to be used by well defined groups.

Random group builder. The random group builder assigns each user to a group randomly, until the desired group size is reached.

Similar users group builder. This strategy is inspired on the work described by Baltruñas et al. [11], which discussed a group formation of any size with the use of the Pearson Correlation Coefficient (PCC) as a measure of similarity. Basically, they define that groups should have a member-to-member high correlation in their preferences. Thus, group members must have a PCC higher than a specified threshold with all other members of the group. However, finding these groups was not an easy task. By its very nature the method cannot assign a group to every user, since not all users in the dataset necessarily have $n - 1$ other users they are similar enough with. As a result, it does not generate overlapping groups and does not cover all users in the dataset.

Our approach to deal with this problem is as follows. We maintain a list of available users (i.e., users that do not belong to any group). Then, the strategy retrieves for each user the other similar users, which are still available, sorted by similarity. An initial group is formed with the firstly selected user as the center. New possible group members are verified for compatibility with the current group. If they have a PCC higher than the predefined threshold with all current members, they are added to the group and the next user is verified. This is performed until the specified group size is reached or all available users have been verified. Once the group reaches the desired size, all its members were removed from the available user list and the process continues with the next available user. If the group do not reach the desired size with the first user as its center, the method will continue with the next available user as center. This method generates groups surrounding an initial user as center and seeks in a greedy fashion the closest group w.r.t. the center. The correlation threshold can be adjusted to generate more or less groups up to a certain point, since groups that have correlations close to 0 are not considered “similar”.

2.3. Group Modeling

Group modeling refers to how individual preferences can be combined to express the group preference, either to rate an item or to generate a ranking. The strategies included in the framework are: *Additive Utilitarian*, *Most Pleasure*, *Least Misery*, *Multiplicative Utilitarian*, *Borda Count*, *Approval Voting*, *Average Without Misery*, *Fairness*, *Plurality Voting* and *Copeland Rule*. All implemented as detailed in [4], except for *Borda Count*.

The *Borda Count* strategy expects that the whole group expresses a rating for all the items that are being considered for the group. This was not always the case in the framework since every member of the groups may have different items for which it is being tested and therefore recommendations for these items are predicted but not for others. To solve this issue, we use a *Partial Voting Borda Count* strategy [18] in which the value of the votes expressed by each member depends on the number of votes and in the order they form. Note that those strategies that produce a rating value can be applied both for ranking and rating recommendation, whereas those that generate ranking cannot be directly mapped to rating predictions in groups. For rating recommendations, both the predicted and expressed preferences were used for the modeling, assuming the

recommendations do not have to be novel for the entire group. However, for ranking strategies, the individual ratings could not always be considered for the formation of the group ranking, since a direct matching between the rating and a ranking for that user was not possible.

2.4. Splitting

The splitting of rating datasets can be performed in different ways. The most popular strategies include a percentage of the ratings expressed, a percentage of the items rated by each user, a percentage of the ratings for each item, a Leave-One-Out or Hold-On, and Folds strategies. All these strategies are currently offered by LibRec and by our extension. In Figure 1, the group recommendation pipeline includes the split of data, which comes from LibRec, and later a group splitting. This is necessary because in group recommender systems additional considerations should be taken into account: 1) *Should the split be independent of the group structures?*, 2) *What means a percentage of the items “rated by the group”?*. These questions remain unclear in the literature.

Regarding the *first question*, we wanted to allow splitting strategies to be either dependent or independent of the groups. Thus, we introduced a new splitting dependent on the groups and maintained the possibility of using the splittings already existing in LibRec, which are independent of the group structures. After the splitting, it is necessary to review test samples. When the group building strategy covers all users in the dataset, the test set is not reviewed. In contrast, others may not cover all users, so the users that do not belong to any group will not get recommendations. In this case, the test set is fixed. That is, test samples that are of users that do not belong to any group are moved back to the train set. By doing this, all users in the test set get group recommendations and can be evaluated. Even though this changes slightly the number of test samples, it only affects group building strategies that do not cover all users.

In relation to our *second question*, Najjar and Wilson [19] defined a group splitting strategy that mimics the rating percentage of users but for groups. We have included it, with a relaxation on the constraints. Instead of considering only items rated by all members in a group, which would not be viable in most datasets, we consider as candidate test item any item rated by a member of a group. All ratings of group members that belong to the excluded items are then moved to the test set. This manner of splitting can be used to test recommender systems for groups that have the restriction of exclusively novel recommendations. This can be applied in situations in which the usage of an item by any member of a group prevents that item for being considered by the group.

2.5. Effectiveness Evaluation for Groups

In order to evaluate the effectiveness for groups of users, we adopted an approach that compares the recommendations generated for the groups against the expressed preferences by each member in the test dataset. Specifically, when considering a rating prediction setting, we compare the ratings of each user in the test dataset against the ratings predicted for the group. In case a ranking is generated, we evaluate the utility of the ranking generated for the group using the information of the test set of the user.

3. Analysis of the GroupLibRec framework

3.1. Setup

For evaluating the framework, we study the performance in *rating* and *ranking* recommendations for groups in the MovieLens1M dataset, which contains 1M ratings, given by 6000 users to 4000 movies. The specific setup of the framework is as follows.

Groups were built using the *random* and the *similar group building* approaches with group sizes of 2, 3, 4, and 8. As group modeling strategies, we considered *Additive Utilitarian*, *Least Misery*, and *Most Pleasure* for the rating experiments, and all the modeling techniques for ranking experiments. The splitting between train and test was performed using the *ratio percentage splitter* included in the LibRec library with 80% of ratings for training and 20% of ratings in the dataset for testing. Groups were generated and stored previously to recommendations predictions. The number of groups generated depended on the size of the groups being tested and the approach. For the *similar group building* strategy 0.27 was used as threshold for the similarity between members.

In addition, we have chosen three recommendation algorithms for individual user predictions, namely *UserKnn* (User K-Nearest Neighbor) and *BiasedMF* (Biased Matrix Factorization) for rating and ranking prediction, and *BPR* (Bayesian Personalized Ranking [20]) solely for ranking. The *BiasedMF* recommender was used with a learning rate of 0.01 and 20 factors. The *UserKnn* recommender used 20 nearest neighbors and PCC as similarity measure. *BPR* was used with a learning rate of 0.1, 10 factors and a decay of 1.0. We use two metrics for the evaluation, *MAE* (*Mean Absolute Error*) is used for the rating prediction results and *NDCG* (*Normalized Discounted Cumulative Gain*) for the ranking predictions. For the ranking experiments, we generated a top-20 for each group.

3.2. Analysis of Results

In this section we analyze the effectiveness of our framework.

Rating prediction. We begin by comparing the effectiveness of each group modeling strategy in the rating prediction task. Figure 2 depicts the results for the different prediction algorithms and group building strategies, in relation to the group modeling strategies analyzed. It is important to highlight that in all cases the *BiasedMF* algorithm outperformed *UserKnn*. Moreover, the performance of the recommendation algorithms was slightly better in the similar groups than in the random ones (see dotted lines). Baltrunas et al. [11] did a similar comparison of group building strategies with Movielens 100k, with results comparable to ours. Indeed, they established 0.27 as threshold for user similarity, because of the distribution among the similarity pairs, and we verified that in Movielens 1M it was the same. In addition, in Figure 2 we compared the rating performance of Additive Utilitarian, Least Misery, and Most Pleasure with respect to the group size. Our results denote that Additive Utilitarian performed the best in spite of the base recommendation algorithm. Another remarkable observation is that independently of the group building strategy, all algorithm's performance decreased as group size increases. Boratto and Carta [21] also noted this effect. The only exception to this being the *Additive Utilitarian* model with the *UserKNN* recommender in the similar groups, which maintained itself almost stable. Indeed, this combination improves with the size of the groups because the bigger the groups are, the more users end up forming part of the

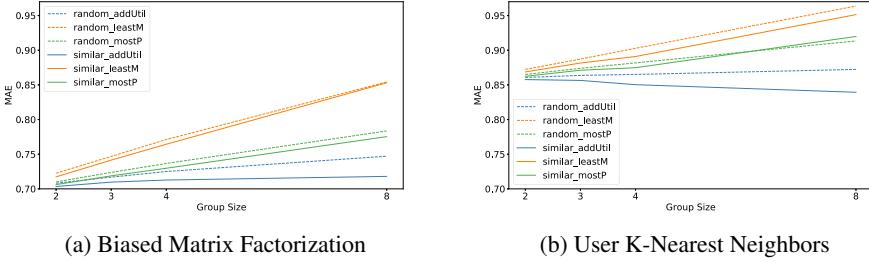


Figure 2. Rating performance comparison by base recommender algorithm. Base single user recommendation algorithm performance in groups of size 2, 3, 4 and 8 for random and similar group building strategies. Performance is being measured with the Mean Absolute Error metric.

neighbors considered for the recommendations. Despite that, the value of the MAE for this configuration is still higher than that of *BiasedMF*.

Ranking generation. In Figure 3 we first analyzed the performance of all group models for each recommendation algorithm separately. Similarly to the rating scenario, the results show that in the random group formation the performance decreases with the size of the groups. This seems an expected behavior since the bigger the group, the less probable it is that the right items for each user will be present in the top 20. However in the similar groups, in some situations, the performance minimally improves or is stable with respect to the group size. This may be due to the fact that similar groups are formed based on their expressed ratings and the bigger the groups, the more items in common all members have and thus increasing the chances of them being present in the group top 20 items.

Execution time. It is also an important factor for a framework, we have compared the time required to execute each modeling strategy in the approaches optimized for the ranking. This choice was made because rating prediction strategies are few in our comparison and they required only a few seconds in the MovieLens 1M dataset, while ranking recommendations take more time due to the number of items being considered for each member in the group ranking process. It is important to highlight that all the strategies obtained good results (from 12 to 50 seconds), with the exception of Copeland Rule. This is largely due to the complexity of this strategy, which needs to consider all the pairwise victories of items in the rankings of each group member. Such number of combinations explodes with the number of items and the number of group members. As a result, the execution of Copeland Rule takes approximately one hour for each experiment, whereas the remaining strategies spend less than a minute.

Discussion. The first aspect that comes out of our evaluation is that we have proved again some of the assumptions already expressed in previous papers regarding group recommendations and group sizes. Indeed, in Figs. 2 and 3, the results show that the smaller the group size, the better the performance. Additionally, we have also shown how the framework allows for reproducibility and quick testing of a great variety of group recommendation algorithms. An important aspect in a framework is the execution time. With the exception of the Copeland Rule group modeling strategy, the results clearly show that the remaining strategies are under reasonable limits, allowing the usage of the framework in small and medium sized datasets. One great advantage of our way of extending the LibRec library is that previous users can simply change configuration files and execute

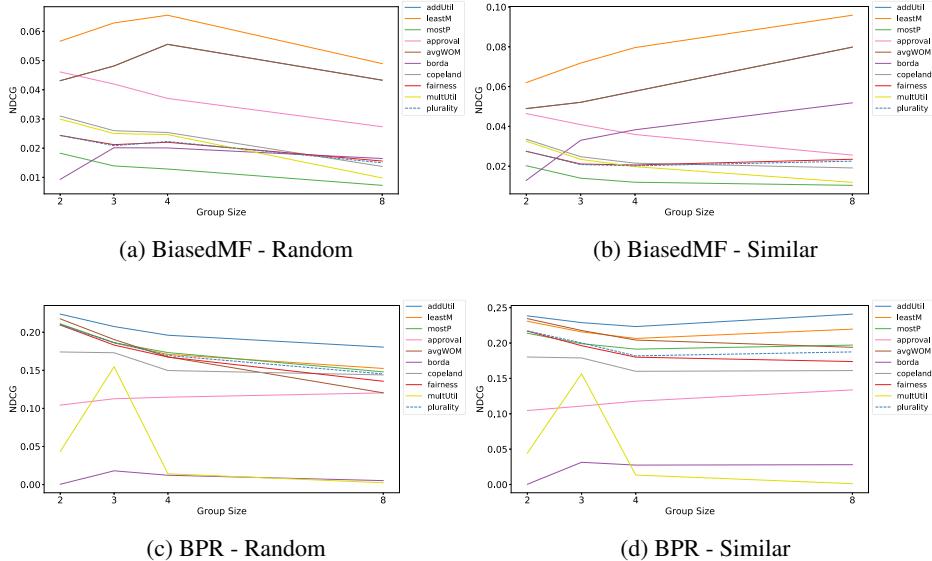


Figure 3. Ranking performance comparison between group modeling strategies. (a) and (c) show the performance in the random group building strategy for the BiasedMF and BPR. (b) and (d) show the same algorithms in the similar group building approach. Performance is measured with NDCG in the top20 rank.

the same experiments used in individual recommendations but for group recommendations. Finally, regarding the limitations of our work, despite allowing for a large number of combinations of strategies, general architectures for building the recommendations other than aggregating individual user predictions are not covered by the framework. For example, the usage of the groups structure for building the individual member recommendations before aggregating them into the group recommendations (such as [19] for memory-based group recommendations). Our expectation is to cover this in the future.

4. Conclusions and Future Work

In this paper, we tackled the issue of enabling reproducibility in group recommender systems due to the lack of frameworks in the field to build, test and benchmark this type of recommenders. We have proposed GroupLibrec, a framework on the scope of collaborative filtering algorithm, that concentrates on three of the main stages of a group recommender system: *group formation*, *group modeling*, and *evaluation*. In any of these three stages, we have integrated several strategies in order to enable the reproducibility of the most used strategies in the field. Indeed, we have shown how the structure of the framework allows for a great number of combinations of base recommendation algorithms, group modeling strategies, and group building strategies. The proposed framework can be easily used for reproducing, testing and helping with the development of new group recommendation systems.

As future work, we plan to include even more features, such as the ability to use other group recommendation architectures or the use of external information in the group building stage.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860843.

References

- [1] Beel J, Breitinger C, Langer S, Lommatsch A, Gipp B. Towards Reproducibility in Recommender Systems Research. *User Modeling and User-Adapted Interaction*. 2016 Mar;26(1):69–101. Available from: <https://doi.org/10.1007/s11257-016-9174-x>. doi:10.1007/s11257-016-9174-x.
- [2] Ferrari Dacrema M, Cremonesi P, Jannach D. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: Proc. of the 13th ACM Conference on Recommender Systems; 2019. p. 101–109. doi:10.1145/3298689.3347058.
- [3] Ricci F, Rokach L, Shapira B. *Recommender Systems Handbook*. 3rd ed. Springer; 2022.
- [4] Boratto L, Felfernig A. Group Recommendations. In: *Collaborative Recommendations - Algorithms, Practical Challenges and Applications*. WorldScientific; 2018. p. 203–232. doi:10.1142/9789813275355_0006.
- [5] Jameson A, Willemse MC, Felfernig A. In: *Individual and Group Decision Making and Recommender Systems*. Springer US; 2022. p. 789–832. doi:10.1007/978-1-0716-2197-4_21.
- [6] Garcia I, Sebastia L, Onaindia E, Guzman C. A Group Recommender System for Tourist Activities. In: Proc. 10th Int. Conf. on E-Commerce and Web Technologies,; 2009. p. 26–37. doi:10.1007/978-3-642-03964-5_4.
- [7] Sánchez LQ, Recio-García JA, Díaz-Agudo B, Jiménez-Díaz G. Social factors in group recommender systems. *ACM Trans Intell Syst Technol*. 2013;4(1):8:1–8:30. doi:10.1145/2414425.2414433.
- [8] Contreras D, Salamó M, Pascual J. A Web-Based Environment to Support Online and Collaborative Group Recommendation Scenarios. *Applied Artificial Intelligence*. 2015;29(5):480–499. doi:10.1080/08839514.2015.1026661.
- [9] Contreras D, Salamó M, Boratto L. Integrating Collaboration and Leadership in Conversational Group Recommender Systems. *ACM Trans Inf Syst*. 2021;39(4). Available from: <https://doi.org/10.1145/3462759>.
- [10] Amer-Yahia S, Roy SB, Chawla A, Das G, Yu C. Group Recommendation: Semantics and Efficiency. *Proceedings of the VLDB Endowment*. 2009;2(1):754–765.
- [11] Baltrunas L, Makcinskas T, Ricci F. Group Recommendations with Rank Aggregation and Collaborative Filtering. In: Proc. 4th ACM Conference on Recommender Systems. RecSys '10. Association for Computing Machinery; 2010. p. 119–126. doi:10.1145/1864708.1864733.
- [12] Salamó M, McCarthy K, Smyth B. Generating Recommendations for Consensus Negotiation in Group Personalization Services. *Personal Ubiquitous Computing*. 2012 Jun;16(5):597–610.
- [13] Recio-García JA, Jiménez-Díaz G, Sánchez-Ruiz-Granados AA, Díaz-Agudo B. Personality aware recommendations to groups. In: Proc. ACM Conference on Recommender Systems. ACM; 2009. p. 325–328. Available from: <https://doi.acm.org/10.1145/1639714.1639779>.
- [14] Jameson A, Smyth B. Recommendation to groups. In: Brusilovsky P, Kobsa A, Nejdl W, editors. *The adaptive web*. Berlin, Heidelberg: Springer-Verlag; 2007. p. 596–627.
- [15] Boratto L, Carta S. State-of-the-Art in Group Recommendation and New Approaches for Automatic Identification of Groups. In: *Information Retrieval and Mining in Distributed Environments*; 2011. p. 1–20. doi:10.1007/978-3-642-16089-9_1.
- [16] McCarthy K, Salamó M, Coyle L, McGinty L, Nixon BSP. CATS: A Synchronous Approach to Collaborative Group Recommendation. In: Proc. of the FLAIRS 2006 Conf. Springer; 2006. p. 1–16. Florida.
- [17] Masthoff J, Delić A. In: *Group Recommender Systems: Beyond Preference Aggregation*. New York, NY: Springer US; 2022. p. 381–420. doi:10.1007/978-1-0716-2197-4_10.
- [18] Koffi C. Exploring a generalized partial Borda count voting system. Senior Projects. 2015.
- [19] Najjar NA, Wilson DC. Differential Neighborhood Selection In Memory-Based Group Recommender Systems. In: Proc. 27th FLAIRS Conf.; 2014. p. 69–74.
- [20] Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L. BPR: Bayesian Personalized Ranking from Implicit Feedback. In: Proc. 25th Conf. on Uncertainty in Artificial Intelligence; 2009. p. 452–461.
- [21] Boratto L, Carta S. ART: group recommendation approaches for automatically detected groups. *Int J Mach Learn Cybern*. 2015;6(6):953–980. doi:10.1007/s13042-015-0371-4.

Contextual TV Show Recommendation

Paula Gómez DURAN¹, Jordi VITRIÀ

Departament de Matemàtiques i Informàtica, Universitat de Barcelona

Abstract.

Recommender systems are a form of artificial intelligence that is used to suggest items to users of digital platforms. They use large data sets to infer models of users' behavior and preferences in order to recommend items that the user may be interested in. Following the trend imposed by digital media companies and willing to adapt to the media consumption habits of their customers, TV broadcasters are starting to realize the potential of recommender systems to personalize the access to their online catalog. By understanding what viewers are watching and what they might like, TV broadcasters can improve the quality of their programming, increase viewership, and attract new viewers.

In this work, we analyze one specific group of users that TV broadcasters must take into account when creating a recommender system: non-logged users. In this scenario the challenge is to use contextual information about the interaction in order to predict recommendations, as it is not feasible to use any kind of information about the user. We propose a method to leverage data from other type of users (logged users and identified devices) by using Graph Convolutional Networks in order to come up with a more accurate recommender system for unidentified users.

Keywords. Recommender Systems, Graph Convolutional Network, Context-aware recommendations, Public Media Service

1. Introduction

Recommender systems are a family of artificial intelligence tools that are used to suggest items for users of digital platforms, such as online media (OM) platforms and public-service media (PSM) platforms. In this scenario, they are specifically used to optimize user's engagement. Collaborative filtering has been identified as one of the best technical approaches to accomplish this task [3, 5, 19], but the case of PSM platforms faces some specific requirements, such as the presence of a special group of users that are not present in OM and which we aim to analyse in this work: non-identified users. This kind of users is very typical of PSMs, as on most of these platforms there is no requirement for a user to be registered and cookies cannot always be guaranteed to track users' sessions.

Over the last recent years, context-aware recommendations [11] have aroused in order to end up with more accurate recommendations. Context-aware recommendations [1] take into account all aspects surrounding a user's situation when making suggestions.

¹Corresponding Author: Paula Gómez Duran, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain; E-mail: paula.gomez@ub.edu.

This can include factors such as time, location, device, weather, previous social media interactions, and other activities taking place at the same time. However, a lot of this research have been focused on the conventional context-aware problem, in which the model includes context as an important, but not essential, feature in order to enrich the information about the user-item interaction [8, 14]. Hence, those cases where there are no recorded user interactions (continuous cold-start problem) and where context-state is set as the only input (and thus it becomes essential), are under-represented scenarios in recommender systems research.

In this work we state that, in spite of the fact that PMS cannot always guarantee very specific contexts, pure-contextual recommendation problem can be addressed in a positive way. In this scenario we propose to leverage collaborative information from other users by using Graph Convolutional Networks to generate richer predictions than not just those corresponding to the most probable items for a given context specification. Therefore, our only approach is to use a pure contextual recommender system in order to predict recommendations, as it is not feasible to use any other kind of information. By working with a dataset of real interactions from a Catalan PMS, we will show how to build a pure contextual recommender system that maximizes the hit rate and leverage all the previous collaborative data available, even not needing it to have any information about the existing user or items.

2. Related work

Pagano et al. [11] were among the first to identify the huge importance of context in recommenders. They stated that context should be taken into account *per se* instead of being used as an "extra" feature that could only help to get better predictions. They claimed we should move from context-aware to context-driven recommendation models.

Based on the work of Adomavicius and Tuzhilin [1], we can consider that there are two different views of context: representational and interactional. The first one refers to observable attributes that are known a priori and that do not change over time, while the second one refers to the cyclical relationship between context and activity that makes the context change over time. They reckon that fully observable factors are easier to handle than unobservable ones and they present three different ways of including context into a traditional recommender system model: pre-filtering, post-filtering or modeling. In an experimental evaluation, Panniello et al. [12] found that the optimal choice of a pre-filtering or post-filtering strategy strongly depends on the particular recommendation problem. Thus, in this work we will address the context with the modeling strategy.

Context-aware recommendation has become increasingly sophisticated in the recent years [8, 14, 15]. On the one hand, lots of efforts have been made to model *sessions*, which are basically sequences of user interactions, as contextual signals. As a session has usually an aim, there are many different recommender systems that characterize sessions in order to provide better recommendations. On the other hand, using contextual factors in order to figure out the availability patterns of some items have also been widely researched. However, most of the research has been focused on session-based or context-aware models where at least some information of the user is still available (i.e. ID from logged user, session of the user, cookies of the user, ...) while not big efforts have been put in those extreme cases where user and item dynamics make the challenge be a continuous cold start problem.

Since the first models, which simply considered item/context combinations as inputs of the recommender (i.e., a TV program was transformed into 2 different items when it was consumed from TV or it was consumed from a desktop), some more sophisticated models [9, 10, 18] have aroused which open up new possibilities for exploiting new context-variables which help improve the models. However, they require some specific information such as content descriptions or location information to, at least, encode somehow the preferences of a given recommendation (like "simulating" user's ID).

Our work differs from the SOTA methods as we aim to solve the recommendation problem in a more generic way for one specific group of users that TV broadcasters must take into account when creating a recommender system: non-logged users.

3. Logged users' behavior as a context

In this section, we briefly describe the dataset from the Catalan PMS that we have analyzed in the experiments and the method we propose to leverage the information about logged users.

3.1. Dataset

We obtained an anonymized dataset corresponding to historical data of user views of the online catalogue of a Catalan public broadcaster, TV3, collected over diverse platforms (SmartTVs, website, mobile apps) throughout a whole calendar year (2021). The raw data contained information on user interactions indicating *userID* (including fake IDs for non-logged users), *itemID* (a one hot encoding with a '1' activating the position which references the item number) and some contextual information of the interaction such as device or timestamp (using multi-hot encoding to represent the active context combinations). Items were identified at the single episode level (e.g., morning and evening news had separate IDs, as had each episode of TV series as well). As it would not be helpful for the recommender to have to rank different episodes of the same TV show, we aggregated all episodes from the same program into the same *itemID*. This resulted in a large reduction in the number of items.

We identified three groups of users: those who were logged into the platform (logged users), those who we could track by the cookies (cookies users) and those who had just one interaction, either because they did just interact once or due to the impossibility of tracking them (non-logged users).

We report in table 1 the statistics from the non-logged users, which we aim to analyse, and from the logged users dataset, which we have used in order to leverage the past behaviour and the data flow structure that actually occurs in the PMS platform.

	#users	#items	#interactions
Logged	10,919	509	69,161
Non-Logged	251,561	613	15,590,592

Table 1. Dataset statistics in terms of number of users, number of items and number of interactions among them.

The first subgroup of the data is conformed by those users who have a profile on the broadcaster platform. Even though no user information is collected, the fact that they

are logged in is very useful in order to ensure continuity of the data consumption and leverage all user records.

We applied some filters in order to consider data corresponding to a minimum of three interactions per user and also a minimum frequency of three visualizations per item, with the aim of removing outliers and work with more stable data. Due to the diverse nature of the platforms where items are offered, the public TV broadcaster does not require to be logged in and so that is the reason why the dataset is smaller in terms of the number of users (reflected in the summary statistics in table 1).

After filtering, we adapted the data to build a dataset which consist of $\langle userID, itemID, rating, timestamp \rangle$ tuple interactions.

The second subgroup collected is conformed by those users who had just one interaction captured in the platform. In order to determine which are those context that are relevant when trying to characterize an item consumption, we have done some user-profiling analysis by applying some effective graph-based clustering methods such as Markov clustering algorithm or more classic clustering methods such as K-Means. By applying them we could analyze which context dimensions where the ones affecting each cluster and, in fact, we concluded the dimensions that actually affect a user: the device from which the consumption is done, the period of the day (late night, early morning, morning, noon, evening, night), and whether it is weekend or not.

After analysing the most representative context, we adapted the data to build a dataset which consist of $\langle itemID, rating, deviceID, periodID, isweekend \rangle$ tuple interactions, such as for example $\langle 237, 1, 3, 4, 0 \rangle$ would mean that item 237 was consumed ($rating = 1$) in a mobile phone ($deviceID = 3$), during the early morning ($periodID = 4$) of a weekday ($isweekend = 0$).

3.2. Pure-contextual recommendation

As it has been mentioned in previous sections, the most commonly addressed problems in contextual recommendation are those stating that users' preferences could be different across different contextual situations. However, very few research has been done into finding models which suit best for those scenarios where no information from user is collected and in which, therefore, an extreme cold-start problem scenario is created. At that stage, if no information is collected from user nor any information from content can be leveraged (PMS do not always characterize all content they have), the only option left is to build a pure-contextual recommender by modifying some existing models in order to allow them dropping the $userID$ dimension.

Pagano et al. [11] propose some models where $userID$ is just dropped. However, it is non-trivial how to do it, as it would imply changes when evaluating the system or performing negative sampling when training it. Besides, unless the number of unique contexts you have is on the order of number of users in your recommendation data, in which case you would be able to consider each unique context combination as a particular user who consumed an item, a vanilla RS will not be able to succeed this way. In fact, for those cases where the number of unique context is on much minor order than the number of items (e.g. maybe 10, 20, 30 contexts vs 600 items), addressing the problem as a multi-class classification task might be the best option.

In this work, we address the case where, given a unique combination of contexts (which are not enough precise to be used as an identifier for a given interaction), a classi-

fier is able to predict the probability of each item (multi-class problem) to be consumed. In other words, we try to find out a probability for consuming an item under a specific context combination.

3.3. How to leverage historical data from logged users?

As the scenario we are approaching requires to use a **pure contextual recommender system** in order to predict items, we have chosen to treat it as a classification problem, where the items on the database are represented in the output of the network in terms of a probability distribution.

The ideal scene would be to have as many content information as we could about the items to enrich the classifier, as besides contextual information, specifically in those scenarios where unique context combinations are on the order of tens, some extra information is needed in order to provide to the classifier more notions about the items and thus avoid recommending lots of items with the same probability under a given context. However, for some PMS and specifically in the case we show, most of this information have not been consolidated in the same format yet, it is not accessible for legal reasons or sometimes it is not even linked to the item id in consume data (they might be used in the platform for different purposes and sometimes they are not in charge of the same section on the company).

3.3.1. Graph Convolutional Embeddings

In front of this situation, we propose to use the collaborative data available from **logged-users**, fit into a Graph Convolutional Neural Network (GCN), in order to leverage the bias that this kind of models capture and introduce it into the classification model. This way, we would be able to enrich the pure-contextual recommender just in a straightforward manner by leveraging the inductive bias that those models supply.

Specifically, this approach leverages the work proposed by Duran et al. [4] to use Graph Convolutional Embeddings in collaborative recommender systems. We created a *bi-parted graph* by selecting $\{userID - itemID\}$ from the **logged-users** interactions in order to build an adjacency matrix (A). Then, we fit this matrix A into a Graph Convolutional Embedding Layer in order to generate the item embeddings. After performing the convolution, those embeddings will be connected implicitly by user's nodes in such a way that the resulting model is biased by the graph neighborhood structure: items that are similar will have similar embeddings. Of course, this measure of similarity is what we want to take from the **logged-users** data of the OM.

To perform the embeddings we use the following equation:

$$E = \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} H W \quad (1)$$

where E are the resulting items' embeddings, A is the adjacency matrix of the graph that we have just built from **logged-users**, H is the identity matrix (because for the first layer should be the node's features and we do not have content information for any node) and W are the weights we want to learn. Note that in GCN, the D is the degree matrix computed from the \hat{A} , which is just the A matrix with the self-connections added.

However, as the information that we want to propagate is the one referring to the item nodes, in order to avoid loosing information as we train the model and specifically for those items with few connections, we apply some modifications to the initial **GCE layer** and we end up with [Equation 2](#).

The major component of graph convolution is the neighborhood message passing, which allows the supervision signal to flow in the graph and propagate through the edge. Recent works [7, 17] have demonstrated that stacking proper layers of convolution may lead to better performance, especially when the supervision signal is sparse. In RS data, the edges between nodes in the constructed graph are relatively sparse, as a result a single layer of convolution is not enough for sufficient information propagation. However, simply stacking more convolution layers would also increase the risk of over-fitting because the number of trainable parameters would also increase. To address the problem, we utilize a “batched” operation in one single layer to perform multi-hop information propagation with only one parameter matrix. We define the embedding matrix after k -hop propagation as:

$$E = (\hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}})^k HW \quad (2)$$

This enables the node to get information from k -th order neighbourhood with the same number of parameters as a single naive graph convolution layer. Thus, it helps to alleviate the data sparsity problem of recommendation scenarios, specially in this case where the user nodes which connect several items are never actively trained.

3.4. Embedding classifier

The next step is to use these graph-based item embeddings to built a better context-based classifier. Our hypothesis is that by enriching the context information with a GCE embedding layer, we will allow the model to leverage the intrinsic bias that the original database actually holds.

To this end, we propose a new context-based item classifier architecture, called **embedding classifier**, which is represented in [Figure 1](#). On the left side of the figure, it can be seen that the model takes as an input the one-hot encoding of each context in order to compute its respective embeddings and concatenate them, thus ending up with a matrix containing a context embedding in each row. At the same time, on the right side of the figure, the model will take as an input the graph embeddings of each of the items in the dataset, thus ending up with a matrix that will contain in each row an item embedding. Depending on whether we use GCE layer or vanilla embedding layer in order to compute the item embeddings, we will leverage the historical information (or not), respectively.

Then, each of the two matrices (the matrix containing the context embeddings and the matrix containing the item embeddings) are multiplied hence giving a vector of item probabilities for each given context. Finally, those vectors are fed into a linear layer so that the probability of each given item along different contexts is combined in order to end up with a final representation that will represent the probability of an item being consumed under a given combination of contexts.

To sum up, we propose a model that can be trained end-to-end by and which leverage all the previous collaborative data available (no content information is used) just

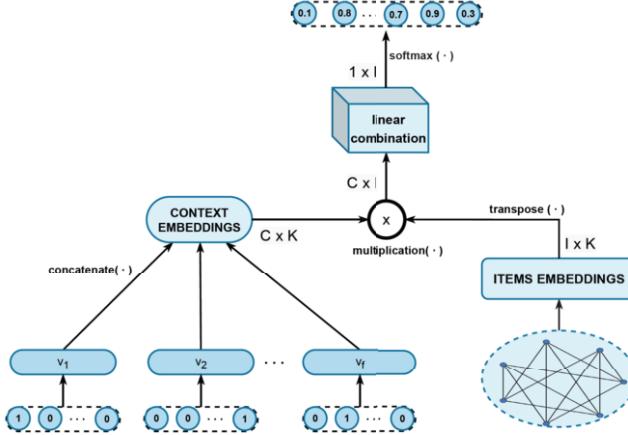


Figure 1. Schema providing an intuition of where GCE layer is applied into the **Embedding Classifier**, as explained in 4.1.

by previously building the A matrix which encode the bias contained into the *user-item* topology from the *logged – users* data. So, the final model is a classifier where the output is a tensor containing the probability assigned to each item of being consumed under a given context combination.

4. Results

In this section we explain in detail the baseline models and the metrics we use in order to reach conclusions for the Catalan PMS dataset (TV3) in contextual recommendation. All the baseline models are state-of-the-art models we chose as realistic models that actually are used or suitable to be used by a public service broadcaster. Then, we compare each of the models with our proposed method (classifier + GCE). Note that we require the chosen models to be capable of being naturally modified to perform pure context-aware recommendations, which moreover implies in this case being treated as a classification problem. That is mainly the reason why we do not choose any recurrent model, as they would require some temporal information of the user consumption's to simulate a sequence that we actually do not have.

4.1. Baselines

We have chosen this baselines, which are used very often in PMS production scenarios for pure-contextual recommendations.

1. **Random:** This model is insensitive to data distribution. It uniformly recommends items independently of the context.
2. **MostPop:** MostPop stands for Most Popular items recommender. This model ignores most of the items and always recommends the K most popular items (the items that were consumed the most), independently of the context.

3. **MLP:** MLP stands for Multi-layer Perceptron classifier, which consists of three layers of units: an input layer representing the context, a hidden layer, and an output layer with one unit for each item to be recommended.
4. **Embedding Classifier (CEmb):** This model we propose is inspired in Factorization Machines [13] with the aim of using embedding to characterize each item and each context option.

4.2. Parameter settings

To ensure a fair comparison of the models' performance, we train all of them by optimizing the Cross Entropy (CE) loss as the training function² with the Adam optimizer[6]. We use Bayesian Optimization[2] strategy in order to tune the hyperparameters and follow the same criteria in all cases. Thus, we determine the best hyperparameters for each of the models by tuning the learning rate on the range {0.0001, 0.0005, 0.001, 0.005, 0.01}; batch size on the range {256, 512, 1024, 2048, 4096}, and dropout on range {0, 0.15, 0.3, 0.5}. The embedding size is set to 32 for all models. We run all the experiments for a maximum of 150 epochs and perform early stopping when the loss stops decreasing for more than 10 consecutive epochs.

4.3. Evaluation protocol

When evaluating a recommender system, the goal is to generate a ranked list of k items by decreasing predicted score of how likely a user is to interact with them. We split 80% of the interactions for training set and 20% for test set. In fact, we sort by timestamp and left out for testing the last 20%, thus simulating the past and future times.

The performance of each model follows from assessing the recommendation lists provided for all users through a range of metrics. In terms of measuring how good a model is, we used standard offline top- k metrics:

- **Hit Rate (HR@ k):** A recall-based metric, measuring whether the test item is in the top- k positions of the recommendation list provided by the RS to a user, averaged across all users.
- **Normalized Discounted Cumulative Gain (nDCG@ k):** It is a rank-sensitive measure of quality of recommendations, which provides information on the ranking position of the ground truth sample; higher scores are assigned to the top-ranked items and can be regarded as a weighted version of HR@ k . This metric can be computed as an average across all users [16].

In addition, we used one metric to assess the diversity of items offered to users across the item catalogue (**coverage**), as it is very important for PMS to be capable of guaranteeing item diversity and personalization instead of just sticking to recommend popular content

- **Coverage (Cov@ k):** is the fraction of items that the RS includes as recommendations across all users on the top k recommendations. A RS that consistently recommended a few very popular items to all users would have a small coverage (i.e. ItemPop model), while a RS that is able to recommend the whole item catalogue to users will have Cov@ k = 1 (i.e. random model).

²<https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html>

4.4. Performance

In this sub-section we show the results for all different models in terms of the defined metrics in the previous subsection.

Non-Logged dataset			
	HR	nDCG	Coverage
Random	0.020	0.009	1.000
MostPop	0.039	0.017	0.018
MLP	0.041	0.019	0.061
CEmb	0.042	0.019	0.062
CEmb - GCE	0.085	0.041	0.058

Table 2. Off-line metrics on Catalan PMS test dataset. Acronyms are the same as in Section 4.3.

In the first two rows we show the results of two opposite models that can always be applied when approaching a pure cold-start scenario: Random and MostPop. On the one hand, we show that the random model systematically offers the lowest metrics in terms of succeeding (HR and NDCG), while offering a coverage of 100%, as it goes through all items of the database to offer recommendations. On the other hand, we see that recommending the most popular items leads in general terms to a very good performance. However, this generally means to reduce the number of items from database showed (coverage) to very small number. In fact, we can see that even the performance metrics (HR, NDCG) are high compared with the model complexity, the coverage is reduced to showing just the 1.8 % items from the database.

As we have mentioned in the previous sections, the trade-off between engaging the user to the platform and offering a wide variety of items to the users in order to ease the access to all media type of content to consumers, it really something to consider for PMS. Thus, the best model should not just have very good metrics in terms of performance (as could be MostPop model in some cases) but also try to enlarge coverage to ensure showing as much items as they can from the database without being randomly shown.

In the second and third row, we show two SOTA classifiers that we have chosen: one that do not use embeddings to characterize entities (each context and item) - MLP - and one that, in fact, does - CEmb. We observe in Table 2 that both models coexist in the same range of metrics, having the model which use embeddings to characterize entities a slightly better performance in terms of HR and coverage.

As we have explained, we have chosen the embeddings' classifier (CEmb) model as it is constituted by embeddings that characterize each entity (each context or item). By choosing a model that uses embeddings, we are able to substitute the vanilla embedding layer by the GCE layer and leverage past collaborative interactions. That is the model that we see in the last row, with the one we show that leveraging past information of logged users outperforms all other models in terms of HR and NDCG metrics for more than double. Besides, the coverage of items showed to users in recommendation is just slightly reduced, and hence we end up with a good trade-off that accomplish the policy that follow a PMS.

5. Conclusions

In this paper we have focused on one specific group of users that TV broadcasters must take into account when creating a recommender system: non-logged users. Not a lot of efforts have been put into facing this type of problem, which is basically building a pure-contextual recommender system that end up working in a scenario in which all data follows the cold-start problem distribution: no information at all about the user, session, etc. Thus, we approach the challenge of predicting recommendations for those type of users by using contextual information about the interaction and seeking to leverage data from those logged users, who we state that implicitly hold a bias that describe how data is willing to be consumed.

In the end, by working with a dataset of real interactions from a public Catalan TV broadcaster, we show in the results how building a contextual recommender system that maximizes the hit rate and leverage all data available can actually be accomplished in order to come up with better recommendations for unidentified users.

6. Acknowledgements

This research was partially funded by the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ) under project PICAE (Comunitat RIS3Cat Media). P.G. and J.V. acknowledge funding from projects Nos. RTI2018-095232-B-C21 (MINECO/FEDER, UE) and 2017SGR1742 (Generalitat de Catalunya).

References

- [1] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.
- [2] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015. .
- [3] Xuegang Chen, Mingna Xia, Jieren Cheng, Xiangyan Tang, and Jialu Zhang. Trend prediction of internet public opinion based on collaborative filtering. In *2016 12th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*, pages 583–588. IEEE, 2016.
- [4] Paula G. Duran, Alexandros Karatzoglou, Jordi Vitrià, Xin Xin, and Ioannis Arapakis. Graph convolutional embeddings for recommender systems. *IEEE Access*, 9: 100173–100184, 2021. ISSN 21693536. . URL <http://arxiv.org/abs/2103.03587>.
- [5] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 199–206, 2010.
- [6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv e-print*, 2014.
- [7] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

- [8] Miklas Strøm Kristoffersen, Sven Ewan Shepstone, and Zheng-Hua Tan. The importance of context when recommending tv content: Dataset and algorithms. *IEEE Transactions on Multimedia*, 22(6):1531–1541, 2019.
- [9] Augusto Q Macedo, Leandro B Marinho, and Rodrygo LT Santos. Context-aware event recommendation in event-based social networks. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 123–130, 2015.
- [10] Ante Odić, Marko Tkalčič, Jurij F Tasič, and Andrej Košir. Predicting and detecting the relevant contextual information in a movie-recommender system. *Interacting with Computers*, 25(1):74–90, 2013.
- [11] Roberto Pagano, Paolo Cremonesi, Martha Larson, Balázs Hidasi, Domonkos Tikk, Alexandros Karatzoglou, and Massimo Quadrana. The contextual turn: From context-aware to context-driven recommender systems. In *Proceedings of the 10th ACM conference on recommender systems*, pages 249–252, 2016.
- [12] Umberto Panniello, Alexander Tuzhilin, Michele Gorgoglion, Cosimo Palmisano, and Anto Pedone. Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 265–268, 2009.
- [13] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2010. .
- [14] Nícollas Silva, Heitor Werneck, Thiago Silva, Adriano CM Pereira, and Leonardo Rocha. A contextual approach to improve the user’s experience in interactive recommendation systems. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 89–96, 2021.
- [15] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38, 2021.
- [16] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v30/Wang13.html>.
- [17] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. *arXiv preprint arXiv:1902.07153*, 2019.
- [18] Jiancan Wu, Xiangnan He, Xiang Wang, Qifan Wang, Weijian Chen, Jianxun Lian, and Xing Xie. Graph convolution machine for context-aware recommender system. *Frontiers of Computer Science*, 16(6):1–12, 2022.
- [19] Jinfeng Zhong and Elsa Negre. Towards improving user-recommender systems interactions. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pages 816–820, 2022. .

This page intentionally left blank

Machine Learning

This page intentionally left blank

Predicting Personalized Quality of Life of an Intellectually Disabled Person Utilizing Machine Learning

Gaurav Kumar Yadav ^{a,d,1}, Benigno Moreno Vidales ^b, and Sara Dueñas ^b and Mohamed Abdel-Nasser ^c and Hatem A Rashwan ^a and Domenec Puig ^a and G. C. Nandi ^d

^a *Universitat Rovira I Virgili, Tarragona, Spain*

^b *Instituto de la Robótica Para la Dependencia, Sitges, Barcelona, Spain*

^c *Department of Electrical Engineering, Aswan University, Egypt*

^d *India Institute of Information Technology Allahabad, India*

Abstract. This work aims to enhance dependent persons' quality of life (QOL) by examining various aspects of their lives and providing the required assistance to enhance each aspect of their QOL. We employ machine learning methods to evaluate the eight aspects of QOL and forecast the corresponding index value. Machine learning algorithms input eight aspects of QOL and predict the QOL index value. The QOL Index value says the requirement of the support to a person, and it depends on eight aspects of the QOL. We use our dataset to train the machine learning model. Dataset is collected using the GENCAT scale tool, which takes 69 items and provides the score value for each aspect of the QOL. We apply many linear and nonlinear machine learning regression algorithms. The multiple linear regression algorithm results show better performance for root mean squared error (1.4729) and R^2 score (0.9918).

Keywords. Quality of Life, Intellectual and Developmental Disability, Priority of Care, Support Paradigm, and Machine learning.

1. Introduction

The perspective of society has been changed in the current decades toward the rights of intellectually disabled (ID) people. This situation became possible due to adopting the international convention on the right of a person with a disability (CRPD) in 2006. One hundred eighty-five countries ratified it [1] including Spain in 2008, showing their commitment to providing dignity, equality, and freedom to the dependent people. CRPD contains 50 articles, out of the twenty-six, obligate the state to provide legal rights to dependent people in every aspect of life, including social, personal, judicial, Etc. [2]. This work is motivated to improve the quality of life of elderly and intellectually disabled people. We are developing a quality of life support model (QOLSM) that combines the aspects of QOL to support and show the method to implement this concept. With this

¹E-mail: gauravkumar.yadav@urv.cat

motivation, we started to effectuate this idea to provide uninterrupted support to intellectually disabled people using the current development of machine learning techniques. We interviewed 26 beneficiaries individually and recorded their response on a four-point frequency scale. Beneficiaries were asked 69 questions that covered every necessary aspect of people with ID. We used the GENCAT scale [3] tool to convert these 69 question responses into eight aspects of QOL score value and an index value corresponding to the eight aspects value. The paper has the following contributions:

- We apply current machine learning technologies, analyze the eight aspects of QOL of intellectually impaired people and forecast the need for help.
- Using our own recorded dataset to train a machine learning model and provide a method to assist ID personnel.
- Proposed a machine learning-based model to predict person needs support or not to improve their QOL.

2. Methodology

2.1. Dataset

We have the Newton-One dataset, a private dataset collected by us. This dataset contains data of 26 beneficiaries, of which 14 are female and 12 are male. The age range of these recipients is from 65 to 90 years old. This dataset is compiled in the year 2021. It has eight aspects value and corresponding index values, so the original shape of the dataset is 26,9. The eight aspects of the QOL are Emotional Well-being, Personal Development, Physical well-being, Self-determination, Interpersonal relation, Social Inclusion, Material well-being, and Rights. The dataset contains a QOL Index value corresponding to the eight aspects value for each beneficiary. For collecting this dataset, a professional asked 69 questions during the interview of each beneficiary. They recorded the response to these questions on a four-point frequency scale. Further, these questionnaires answer inputting to the GENCAT scale, giving the eight aspects value and corresponding Index value for support. The value of each aspect of QOL varies between 68 to 130.

2.2. ML Techniques

In the Newton-One dataset, the input consists of eight dimensions of quality of life, and the output is the associated index value. Output is continuously dependent on the eight aspects of the QOL, so it is a regression task; therefore, regression algorithms perform well on this dataset. Initially, we do not know the nature of the dataset, so we use both linear and nonlinear regression algorithms. We use multiple linear algorithms, support vector regressor, decision tree, random forest, and gradient boosting algorithms. We have imported these models from scikit-learn and trained them using our dataset.

2.3. Steps of the Proposed Method

Figure 1 shows the fundamental steps of our work. We started our work by collecting a dataset. Professionals interviewed the intellectually disabled people and asked sixty-nine questions, covering all the eight aspects of the QOL of a dependent person. We record

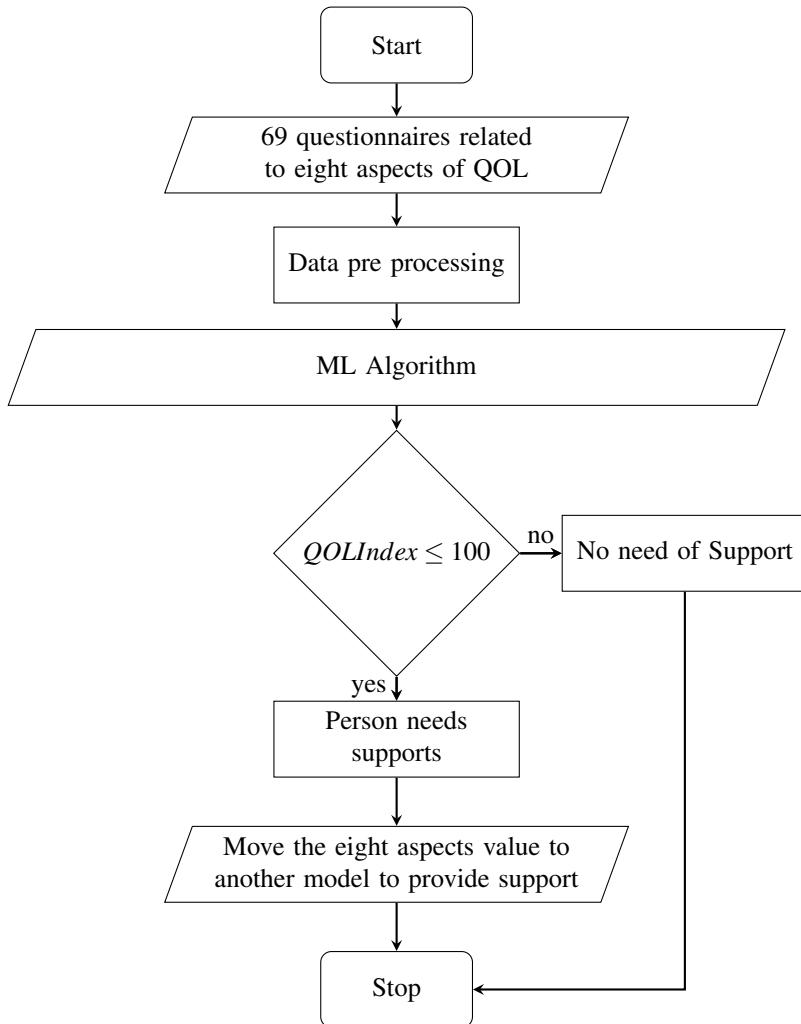


Figure 1. Flowchart of complete architecture of our Machine Learning Model

the answer on four points frequency scale. We build a tabular data which consists of eight columns corresponding to eight aspects of QOL and one column consisting of the Index value. We interview a total of twenty six-person. So the size of the dataset is 26, 9. Each score value of eight aspects and index value varies between sixty-eight to one hundred thirty. Where sixty-eight represent a poor score value, and one hundred thirty shows the best score value. The number of samples in the original dataset is only twenty-six, which is less to train the machine learning models. Therefore we use SMOTE-R [4] algorithm to augment our underfit data. Before augmenting the data, we split data into the train and test case because we want to test our trained model on original test data to validate our model. We split our dataset 80% (20) for training and 20% (6) for the testing set. Then after we augmented the training dataset from twenty to hundred forty-five, we trained each machine learning algorithm and calculated the mean absolute error, root mean square error value, and R^2 score value for train and test cases.

Table 1. Measuring matrices showing the performance of machine learning models trained on an augmented dataset for training and results for original test set dataset

Algorithms	MLR	DT	RF	GB	SVR
Matrices	Train Test	Train Test	Train Test	Train Test	Train Test
MAE	0.4902 1.3504	0.0000 8.5000	2.2340 8.4383	0.1161 6.0268	0.1000 18.7400
RMSE	0.6350 1.4729	0.0000 10.8857	2.6883 9.1498	0.1419 7.4630	0.1000 19.9124
R^2 Score	0.9981 0.9918	1.0000 0.5537	0.9676 0.6847	0.9999 0.7902	0.99995 -0.4931

3. Results

We present the findings as evaluation matrices. Table 1 displays the computed performance for the train and test instance. It shows the result for multiple linear regression algorithms (MLR), decision trees (DT), random forests (RF), gradient boosting (GB), and support vector regression (SVR) algorithms. The outcomes of the MLR algorithm are superior to those of other algorithms in terms of R^2 score value and root mean squared error value. The cause behind the MLR algorithm is that the data set is linearly dependent, and other algorithms are best for nonlinear datasets. In order to anticipate the QOL index value during the test case and for subsequent use, we finalized the MLR algorithm based on the RMSE value and R^2 score. By considering eight aspects of QOL, the MLR algorithm forecasts the value of the QOL Index. Finally, determine whether the person requires support based on the projected QOL Index value model.

It should be noted that the prediction results could be further improved by employing a metaheuristics algorithm like the stochastic whale optimization algorithm [5] to optimize the parameters of the regression techniques.

4. Conclusion

In this work, we analyzed the QOL of intellectually disabled people. We trained numerous machine learning regression methods, both linear and nonlinear. In order to estimate the QOL index value, algorithms take eight aspects of QOL. The evaluation matrices score shows that the performance of the MLR is outperforming the others. We finalized the MLR for further use to predict the QOL index value. The QOL Index value determines if a person needs assistance or not.

References

- [1] “Status of ratification interactive dashboard,” <http://indicators.ohchr.org>, 2022.
- [2] L. E. Gómez, M. L. Morán, S. Al-Halabí, C. Swerts, M. Á. Verdugo, and R. L. Schalock, “Quality of life and the international convention on the rights of persons with disabilities: Consensus indicators for assessment,” *Psicothema*, vol. 34, no. 2, pp. 182–191, 2022.
- [3] M. Verdugo, P. Navas, L. Gómez, and R. L. Schalock, “The concept of quality of life and its role in enhancing human rights in the field of intellectual disability,” *Journal of Intellectual Disability Research*, vol. 56, no. 11, pp. 1036–1045, 2012.
- [4] L. Camacho, G. Douzas, and F. Bacao, “Geometric smote for regression,” *Expert Systems with Applications*, p. 116387, 2022.
- [5] F. Mohamed, M. Abdel-Nasser, K. Mahmoud, and S. Kamel, “Economic dispatch using stochastic whale optimization algorithm,” in *2018 International Conference on Innovative Trends in Computer Engineering (ITCE)*. IEEE, 2018, pp. 19–24.

The Assessment of Clustering on Weighted Networks with R Package *clustAnalytics*

Argimiro ARRATIA^{a,1} and Martí RENEDO-MIRAMBELL^a

^aSoft Computing Research Group (SOCO)

at Intelligent Data Science and Artificial Intelligence Research Center

Department of Computer Sciences,

Polytechnical University of Catalonia, Barcelona, Spain.

argimiro@cs.upc.edu marti.renedo@gmail.com

Abstract. We present *clustAnalytics*, an R package available now on CRAN, which provides methods to validate the results of clustering algorithms on unweighted and weighted networks, particularly for the cases where the existence of a community structure is unknown. *clustAnalytics* comprises a set of criteria for assessing the significance and stability of a clustering. To evaluate clusters' significance, *clustAnalytics* provides a set of community scoring functions, and systematically compares their values to those of a suitable null model. For this it employs a switching model to produce randomized graphs with weighted edges. To test for clusters' stability, a non parametric bootstrap method is used, together with similarity metrics derived from information theory and combinatorics. In order to assess the effectiveness of our clustering quality evaluation methods, we provide methods to synthetically generate networks (weighted or not) with a ground truth community structure based on the stochastic block model construction, as well as on a preferential attachment model, the latter producing networks with communities and scale-free degree distribution.

Keywords. clustering, networks, scoring functions, stochastic block model, non parametric bootstrap, R

1. Introduction

Clustering of networks is a popular research field, and a wide variety of algorithms have been proposed over the years. However, determining how meaningful the results are can often be difficult, as well as choosing which algorithm better suits a particular dataset. To help into this assessment and decision of clustering algorithms we have contributed to CRAN the new R package *clustAnalytics* [1], which contains a suite of novel methods to validate the partitions into communities of networks obtained by any given clustering algorithm. In particular, its clustering validation methods focus on two of the most important aspects of cluster assessment: the significance and the stability of the resulting clusters.

¹Corresponding Author: Argimiro Arratia, e-mail: argimiro@cs.upc.edu

To assess the significance of communities structure, *clustAnalytics* has a collection of community scoring functions that measure some topological characteristics of the ground-truth communities, and whose values are compared against those obtained on null models with similar graph properties but without any expectations of a community structure. To evaluate stability, we designed and programmed in *clustAnalytics* a bootstrap technique with perturbations adapted to clustering on graphs. To compare how the clusters of the bootstrapped networks differ from the originals, three cluster similarity measures are provided: the adjusted Rand index, the Variation of Information, and the Reduced Mutual Information.

clustAnalytics handles weighted networks, as well as unweighted, and contains several other functionalities for producing different statistics on a network. It also contains methods for creating synthetic weighted networks based on the stochastic block model construction [2], and the preferential attachment model of Barabasi-Albert [3] that produce examples of ground-truth networks with community structure and degree distribution either binomial or scale-free. The mathematical and algorithmic aspects of *clustAnalytics* is explained in [4].

2. *clustAnalytics*: Examples of usage

First to exhibit the graph randomization procedure programmed in *clustAnalytics*, we apply it to the Zachary's karate club graph, with the default settings (positive weights with no upper bound, which suits this graph):

```
> library(clustAnalytics)
> data(karate, package="igraphdata")
> rewired_karate <- rewireCpp(karate, weight_sel = "max_weight")
> par(mfrow=c(1,2), mai=c(0,0.1,0.3,0.1))
> plot(karate, main="karate")
> plot(rewired_karate, main="rewired_karate")
```

The function `rewireCpp` produces a random version of the original graph by rewiring the edges while keeping the degree distribution constant. The number of iterations is $Q \cdot \#edges = 100 \cdot 78$, where the parameter Q can be set by the user.

Cluster significance and stability. For gauging significance there is an ensemble of scoring functions in `evaluate_significance` which apply simultaneously to each of the clustering produced on a graph by a given list of algorithms. By default the clustering algorithms are Louvain, label propagation and Walktrap, but the function can take any list of clustering algorithms for *igraph* graphs. The function allows for comparison against ground-truth in case this is known. For the karate club graph this is known and we can include it in the analysis

```
#ground truth clusters for karate graph
> karate_gt_clustering <- c(1,1,1,1,1,1,1,1,1,2,1,1,1,1,2,2,1,1,
+                           2,1,2,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2)
> significance_table_karate <- evaluate_significance(karate,
+                                         ground_truth=TRUE,
+                                         gt_clustering=karate_gt_clustering)
> significance_table_karate
```

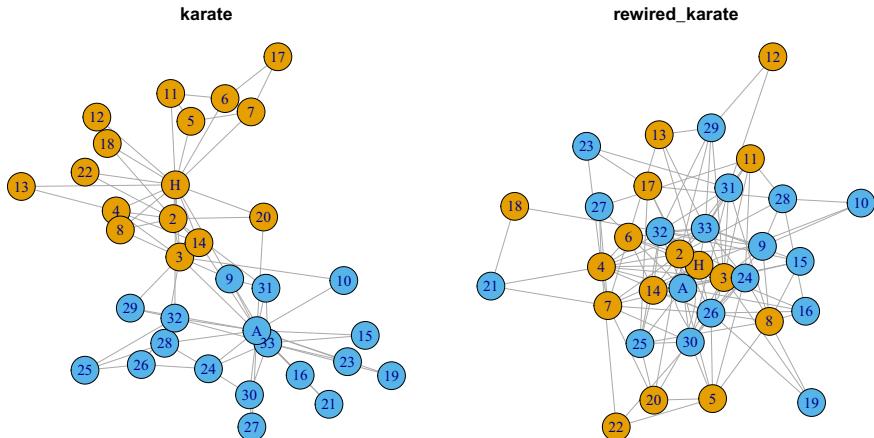


Figure 1. Karate club graph before and after the edge randomization process. Colors represent the fraction of each participant, the *ground truth* clustering in this network.

This prints a table with all scores by the quality measures (Louvain, label propagation and Walktrap) including those for the ground-truth. Now we generate a graph from a stochastic block model in which we set very strong clusters: the elements in the diagonal of the matrix are much larger than the rest, so the probability of intra-cluster edges is much higher than that of inter-cluster edges.

```
> pm <- matrix (c(.3, .001, .001, .003,
                  .001, .2, .005, .002,
                  .001, .005, .2, .001,
                  .003, .002, .001, .3), nrow=4, ncol=4)
> g_sbm <- sample_sbm(100, pref.matrix=pm,
                        block.sizes=c(25,25,25,25))
> E(g_sbm)$weight <- 1
> significance_table_sbm <- evaluate_significance(g_sbm)
> significance_table_sbm
```

We now assess for stability of clustering algorithms. Here we perform a nonparametric bootstrap to the karate club graph and the same selection of algorithms. For each instance, the set of vertices is resampled, the induced graph is obtained by taking the new set of vertices with the induced edges from the original graph, and the clustering algorithms are applied. These results are compared to the induced original clusterings using metrics: the variation of information (VI), normalized reduced mutual information (NRMI) and both adjusted and regular Rand index (Rand and adRand):

```
> b_karate <- boot_alg_list(g=karate, return_data=FALSE, R=99)
> b_karate
      Louvain label prop walktrap
VI        0.2630337 0.2964607 0.2739508
NRMI     0.6957499 0.5447487 0.6698974
```

```
Rand      0.85558310 0.7849259 0.8460001
AdRand    0.6523059  0.5611139 0.6289277
n_clusters 5.6262626 4.9696970 5.8787879
```

And the same for the stochastic block model graph:

```
> b_sbm <- boot_alg_list(g=g_sbm, return_data=FALSE, R=99)
> b_sbm
```

	Louvain	label	prop	Walktrap
VI	0.1234341	0.1769217	0.1178832	
NRMI	0.8536997	0.7841236	0.8656356	
Rand	0.9411244	0.9230160	0.9472768	
AdRand	0.8306925	0.7651778	0.8476909	
n_clusters	6.9797980	7.7070707	7.4646465	

We can clearly see that for all metrics, the results are much more stable, which makes sense because we created the sbm graph with very strong clusters.

Preferential attachment graphs with communities. The `barabasi_albert_blocks` function produces scale-free graphs using extended versions of the Barabasi-Albert model that include a community structure. The parameters that need to be set are m the number of new edges per step, the vector p of label probabilities, the fitness matrix B (with the same dimensions as the length of p), and t_{max} the final graph order. There are two variants of the model. If `type="Hajek"`, new edges are connected with preferential attachment to any existing vertex but using the appropriate values of B as weights. If ‘`type="block_first"`’, new edges are connected first to a community with probability proportional to the values of B , and then a vertex is chosen within that community with regular preferential attachment. In this case, the resulting degree distribution is scale-free (see [5] for a proof of this fact). This is a simple example with just two communities and a graph of order 100 and size 400:

```
> B <- matrix(c(1, 0.2, 0.2, 1), ncol=2)
> G <- barabasi_albert_blocks(m=4, p=c(0.5, 0.5), B=B, t_max=100,
                                type="Hajek",
                                sample_with_replacement = FALSE)
> plot(G, vertex.color=(V(G)$label), vertex.label=NA, vertex.size=10)
```

References

- [1] Renedo-Mirambell M. `clustAnalytics`: Cluster Evaluation on Graphs; 2022. R package version 0.3.1. Available from: <https://CRAN.R-project.org/package=clustAnalytics>.
- [2] Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: First steps. Social networks. 1983;5(2):109-37.
- [3] Barabási AL, Albert R. Emergence of Scaling in Random Networks. Science. 1999;286(5439):509-12. Available from: <https://science.sciencemag.org/content/286/5439/509>.
- [4] Arratia A, Renedo-Mirambell M. Clustering assessment in weighted networks. PeerJ Computer Science. 2021;7:e600.
- [5] Renedo-Mirambell M, Arratia A. Identifying bias in cluster quality metrics. arXiv:2112.06287; 2021. Available from: <https://arxiv.org/abs/2112.06287>.

Binary Delivery Time Classification and Vehicle's Reallocation Based on Car Variants. SEAT: A Case Study

Juan Manuel GARCÍA SÁNCHEZ ^{a,1}, Xavier VILASÍS CARDONA ^a and Alexandre LERMA MARTÍN ^b

^aResearch Group of Data Science for the Digital Society (DS4DS), La Salle-Ramon Llull University
^bSEAT S.A.

Abstract. This note provides a solution to vehicle's compound allocation problem. It has been treated as a classification task employing different Machine Learning (ML) algorithms. It is performed using the known car attributes and the time that vehicles have spent in the compound region, i.e., inventory warehouse, waiting the customer delivery day. Classification results have been assessed with *F1 Score* and *CatBoost* has arisen as the best technique, with values larger than 70%. Finally, reallocation strategy has been tested and outcomes exhibit that company's expert performance is equaled or overcame with respect to time distribution.

Keywords. Machine Learning, Classification, Automotive OEM, F1 Score, Vehicle Reallocation, Anticipatory Shipping, Customer Delivery Time Distribution

1. Introduction

In the last years, automotive Original Equipment Manufacturers (OEMs) are migrating from dealership system to agency model. Both newspapers [1] and consultant companies [2] report about this trend. In this scenario, Machine Learning (ML) can be helpful to automotive OEMs to shipping cars in the region of most likely purchase. This note presents this problem as a binary classification one. The objective consists on **distinguishing whether a car will stay more or less than a threshold days in the compound region, based on the vehicle attributes. Hence, allocate it in the best region.** We prove that ML techniques are helpful to equal and/or improve current delivery time distribution. Compound regions are the equivalent to inventory warehouses managed by the manufacturer.

The article is structured in the following way. Firstly, in Section 2, they are presented related works with the research topic. Hence, Section 3 describes the dataset provided by the automotive OEM source. Next, methodology and results of the research are placed in Section 4 and Section 5, respectively. They are discussed in Section 6. Finally, Section 7 provides conclusions gained and future research paths.

2. Related Works

Mostly of papers focused onto transportation and route optimization. That's why we derived to stock management and product optimization. Reference [3] reviews 49 works

¹Corresponding Author: La Salle-Ramon Llull University, 08024 Barcelona, Spain; E-mail:juanmanuel.g@salle.url.edu

about planning of capacities and build-to-order production. Afterwards, we find a hub of papers supported by demand forecasting. During 2015, researchers of [4] performed a simulation study of an automaker that operates in Brazil by means of demand forecasting and inventory control of spare parts. Recently, in 2021, the work [5] uses on-line retailers' clickstream data and historical sales data to explore the optimal quantity and time of products. Other example is developed for Japanese Seru production system. Authors of [6] are capable of optimizing production quantity allocation, right after optimizing worker allocation problem. Finally, authors in [7] found the correlation between inventory volume and sales in the American automobile market.

Later performing the state-of-the-art review, we discovered a gap in the academia. We did not find evidence of a classification system of vehicles in two types of delivery categories. Especially, one based on car attributes and where categorizing the largest quantity of True Positive class, without neglecting the precision, is a key factor. Therefore, we present this research about allocating the vehicles in the compound region more accurate to reduce the customer delivery time.

3. Dataset description

Data involved in this study is supplied by Spanish car manufacturer SEAT. It collects the time spent by each vehicle from an specific car variant in each compound region within the national market from January 2017 to February 2020, both included. Car variant is defined as the combination of Car Model, Equipment Level (TRIM), Order Type, Exterior Color and Engine. Table 1 explains main descriptive values for each compound region and for the totality of the dataset.

Table 1. Main descriptive values for each compound region individually (Region n) and the whole data (Global) collected in the dataset.

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Global
Min [days]	1	1	1	1	1	1	1
Percentile-25 [days]	26	14	16	14	14	18	15
Percentile-50 [days]	46	29	34	27	25	30	29
Percentile-75 [days]	82	71	82	71	62	69	71
Max [days]	716	447	516	490	470	554	716
Number of Variants	1099	2774	1966	2061	2863	2227	4126
Number of Cars	8670	24526	16608	14216	31874	17432	113326

4. Methodology

The first step consists on assigning classes according to different time thresholds, ranging from 1 to 6 weeks. These binary classes are *Fast Delivery* (FD) and *Normal Delivery* (ND). Correspondingly, the weight of FD class in the dataset varies with the threshold and it is shown in Table 2.

Table 2. Weight of FD (*Fast Delivery*) class over the entire dataset according to threshold time.

	7 days	14 days	21 days	28 days	35 days	42 days
FD cars [%]	5.00	22.67	38.76	48.97	56.05	61.49
FD variants [%]	35.97	65.46	76.30	81.51	84.54	87.03

Secondly, for each threshold, the training process consists on executing cross-validation (5 folds divisions) over data. It is performed under different classification ML algorithms, which are: *Decision Tree*, *Random Forest*, *XGBoost Classifier* and *CatBoost*. This list was defined based on their reliability on other classification problems in the industrial environment [8–10]. With this information, we are able to choose the best algorithm based on *F1 Score*. For the automotive sector, it is relevant to do not only capture as many True Positive as we can, but be precise about the positive class. Papers [11] evidence the employment of this metric in similar contexts.

Afterwards, reallocation step follows for each car variant. In case original region is classified as FD, it is remained. Otherwise, it is headed to all alternative FD destinations. Finally, results are compared with respect to the original situation.

5. Results

Outcomes provided by each ML algorithm can be found in Table 3. The largest values correspond to 28-days threshold. Hence, from this space, *CatBoost* provides the best result. According to Shap values, most relevant features are Order Type and Compound Region.

Table 3. *F1 Score (%)* achieved at each threshold in the cross-validation training process for each ML classification algorithm.

F1 SCORE	7-days	14-days	21-days	28-days	35-days	42-days
Decision Tree	50.05	60.17	69.64	71.13	70.01	68.70
Random Forest	50.43	61.10	70.02	71.32	70.16	68.39
XGBoost	48.74	60.62	71.93	72.75	71.61	69.88
CatBoost	48.72	60.78	72.17	72.90	71.77	70.21

Afterwards, we compare the performance of the reallocation step. We measure the delivery time distribution of all new cars headed to each compound region. These numbers are illustrated on Figure 1.

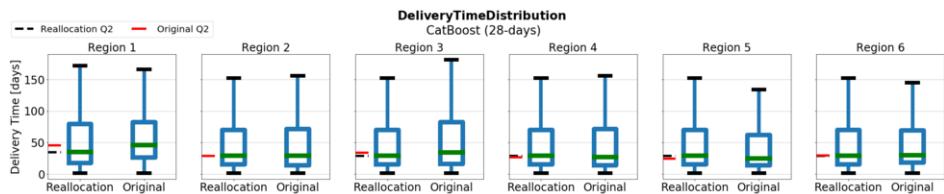


Figure 1. Delivery Time Distribution [days] with and without reallocation for each compound. In each plot of the grid, left boxplot represents Reallocation and right boxplot is Original Delivery Time Distribution. Black dashed line means the median value of the Reallocation Distribution. Red dashdotted line is the median of the Original Distribution.

6. Discussion

From the stage of cross-validation training, the lowest results are given by 7-days threshold for all ML algorithms in Table 3. They reach their peak at the moment of 28-days

threshold, when dataset is almost equally balanced (see Table 2). Regarding reallocation stage from Figure 1, Region 1 and Region 3 are the great benefit from this research. Medians in these compounds are lower than the original situation. In the case of Region 5, reallocation median is slightly larger than the benchmark. For the rest of regions, differences cannot be considered as relevant.

7. Conclusions

Although this note does not include criteria to choose between two or more alternative destinations, nor the capacity of them, it proves that ML techniques are helpful to equal and/or improve delivery time distribution of vehicles. They are based on car attributes, such as Car Model, Order Type, Engine, TRIM, etc. The relevance of the task is crucial in the transition to agency model. We suggest that automotive OEMs use them as supportive tool in the decision making of vehicle allocation.

8. Funding/Acknowledgments

This work is partially funded by the Department de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2019-34.

References

- [1] Ward P. How the agency model is shaking up the car retail industry; 2022. Available from: <https://www.autocar.co.uk/car-news/business-dealership%2C-sales-and-marketing/how-agency-model-shaking-car-retail-industry>.
- [2] Hasenberg JP. Our Automotive Sales News series – Part 2; 2021. Available from: <https://www.rolandberger.com/en/Insights/Publications/How-agency-sales-models-can-benefit-manufacturers-and-dealers.html>.
- [3] Volling T, Matzke A, Grunewald M, Spengler T. Planning of capacities and orders in build-to-order automobile production: A review. European Journal of Operational Research. 2013;224:240-60.
- [4] do Rego JR, de Mesquita MA. Demand forecasting and inventory control: A simulation study on automotive spare parts. International Journal of Production Economics. 2015;161:1-16. Available from: <https://www.sciencedirect.com/science/article/pii/S0925527314003594>.
- [5] Chen C, Xu X, Zou B, Peng H, Li Z. Optimal decision of multiobjective and multiperiod anticipatory shipping under uncertain demand: A data-driven framework. Computers Industrial Engineering. 2021;159:107445.
- [6] Fujita Y, Izui K, Nishiwaki S, Zhang Z, Yin Y. Production Planning Method for Seru Production Systems under Demand Uncertainty. Computers Industrial Engineering. 2021;163:107856.
- [7] Cachon G, Gallino S, Olivares M. Does Adding Inventory Increase Sales? Evidence of a Scarcity Effect in U.S. Automobile Dealerships. Management Science. 2018;65.
- [8] Jabeur SB, Gharib C, Mefteh-Wali S, Arfi WB. CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change. 2021;166:120658. Available from: <https://www.sciencedirect.com/science/article/pii/S0040162521000901>.
- [9] Torgunov D, Trundle P, Campean F, Neagu D, Sherratt A. Vehicle Warranty Claim Prediction from Diagnostic Data Using Classification. In: Ju Z, Yang L, Yang C, Gegov A, Zhou D, editors. Advances in Computational Intelligence Systems. Springer International Publishing; 2020. p. 483-92.
- [10] Wang S, Liu S, Zhang J, Che X, Yuan Y, Wang Z, et al. A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning. Fuel. 2020;282:118848. Available from: <https://www.sciencedirect.com/science/article/pii/S0016236120318445>.
- [11] Zhao S, Li X, Chen YC. A Classification Framework Using Imperfectly Labeled Data for Manufacturing Applications. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). vol. 1; 2020. p. 921-8.

Feature Engineering and Machine Learning Predictive Quality Models for Friction Stir Welding Defect Prediction in Aerospace Applications

Marta CAMPS^{a,1}, Maddi ETXEGARAI^a, Francesc BONADA^a, William LACHENY^b, Sylvain PAULEAU^b and Xavier DOMINGO^a

^a*Eurecat, Centre Tecnològic de Catalunya, Unit of Applied Artificial Intelligence, Av. Carrer de Bilbao, 72, 08005 Barcelona, Spain*

^b*Ariane Group, 51/61 route de Verneuil, Bâtiment 71 - Bureau 142, 78131 – Les Mureaux – France*

Abstract. Data-Driven Predictive Quality solutions are of utmost importance for Industry 4.0 in general and for high added value and complex manufacturing systems in particular. A unique Friction Stir Welding process is performed for the manufacturing of the new Ariane 6 aerospace launchers. This work presents a novel feature engineering approach that correlates Friction Stir Welding process data and quality inspection data to build a Machine Learning-based predictive quality solution. This solution predicts the presence of welding defects, empowering end-user's quality assurance and reducing quality inspection time and associated costs.

Keywords. Machine Learning, Predictive Quality, Industry 4.0, Aerospace Industry

1. Introduction

For product quality assurance, solutions that build upon data analytics, Machine Learning (ML) and Artificial intelligence (AI) can provide major benefits, empowering end-users' decision making that can lead to better products and optimized production lines. H2020 SESAME [1] project aims to bring innovative Industry 4.0 and Data-Driven solutions to the aerospace industry, focusing on reducing the costs of the new Ariane launcher production through data science models for quality prediction, predictive maintenance, and supply chain agility. The aerospace manufacturing industry is highly challenging due to their strict quality requirements and the low production volume.

The use case introduced here studies the Friction Stir Welding (FSW) [2] process on the Ariane 6 Pre-Final Assembly Line. The proposed model, consisting in a set of data analytics and AI/ML-based tools, digests a large amount of welding process data to provide the expert with valuable quality insights that can lead to better and more cost and time-efficient quality control procedures. Following the preliminary study introduced in Camps, et al. [3], this work presents a feature engineering layer as well as the modelling of a binary classifier for the prediction of the presence of the welding defect is presented.

¹ Eurecat, Centre Tecnològic de Catalunya, 08005 Barcelona, Spain; E-mail: marta.camps@eurecat.org

2. Data Fusion & Feature Engineering Layer

For each welding test, the evolution along the 360° (cylindrical tank) of 53 process parameters, including temperature, force, position, and current variables of different FSW subsystems and sensors are acquired together with the FSW station configuration set-up. The sampling frequency of the process parameters varies from one to another (ranging from 0.7 Hz to 330 Hz), resulting in an inconstant timestamping. To obtain a uniform and constant sampling frequency, a resampling methodology based on linear interpolation is implemented.

Once the cylindrical tanks are welded, strategic perpendicular cuts are done to assess the quality of the welding, which is directly related to the penetration of the welding pin into the interface between two welded parts. Even though the process data is acquired for several tests, due to the destructive and expensive nature of the quality inspection, it is performed only in 6 welded tanks and in a few positions of each of those welded tanks. Hence the importance of data-driven solutions that can estimate their values without destructing the welded part. To increment the quality label granularity, linear interpolation is applied.

To get uniform inputs to feed the predictive quality algorithm, a correspondence between process data (timestamped) and quality inspection (space stamped) needs to be implemented. The data is segmented in the same spatial length windows and a set of 9 features are computed for each window of each parameter: the minimum value, the maximum value, the mean value, the standard deviation, the maximum value of the first-order derivative and the first two Power Spectral Density (PSD) computed with Welch's method [4] and their amplitudes. Thus, each window is characterized by 477 new features. To accelerate the computation of the model and to reduce the high correlation observed among some of the features, the Principal Component Analysis (PCA) [5] dimensionality reduction algorithm is applied. Additionally, the two configuration parameters (pin diameter and length) and the Crown variable are added to the input set without the PCA transformation. The input data (process features & configuration parameters) is normalized with the Standard Scaler, which transforms the data to a standard normal distribution with a mean of 0 and a standard deviation of 1 [6].

The process data initially acquired based on the timestamp, and the quality data, based on the position of the quality inspection, are now both referenced to window position, implementing a time-spatial correlation.

3. Model and Hyperparameter Tuning

The proposed predictive quality model is based on a binary classifier algorithm which estimates the quality of the segment, defined as the presence of a defect or not. To explore different strategies and select the best one, different model training strategies determined by the size of the windows or segments, the algorithm and the hyperparameters associated are considered. The exploration focuses in four different windows sizes: 1, 2, 4 and 6 degrees; and six different ML classifiers: Support Vector Machine (SVM) [7], K-Nearest Neighbours (KNN) [8], Decision Trees (DT) [9], Logistic Regression (LR) [10], Naïve Bayes (NB) [11], Bag decision trees (BDT) [9]. For the hyperparameter tuning of each model, an exhaustive grid-search method is implemented, generating candidates from a grid of hyperparameter values specified to evaluate their impact on the model performance.

As there are only 6 tests or files available to train, which are quite different due to the FSW configuration, the training is done by saving one file for the test and using the remaining 5 files to train the model, following a Leave-One-Out strategy [12]. Furthermore, 10 subsets are generated from each training dataset choosing randomly half of the population, thus implementing a dual cross-validation strategy. The best model is chosen by computing the mean of the accuracy metric for each combination of window, algorithm and hyperparameters across all iterations.

4. Results and validation

In this section, the results obtained in the dual cross-validation strategy are presented. Figure 1 shows the accuracy of some of the models generated, for the training in dark blue and test in light blue. Each of the plots in Figure 1 exhibits the results for a given window size and the best hyperparameters for each algorithm. A significant difference is observed between the train and test datasets, especially for 1-degree windows, implying models susceptible to overfitting in this configuration. The best performance is found in the 6-degrees moving window (0.72 for the test), with the smallest difference between train and test accuracy (around 0.28). Therefore, we can ensure that the models with 6-degree windows are more robust and reliable.

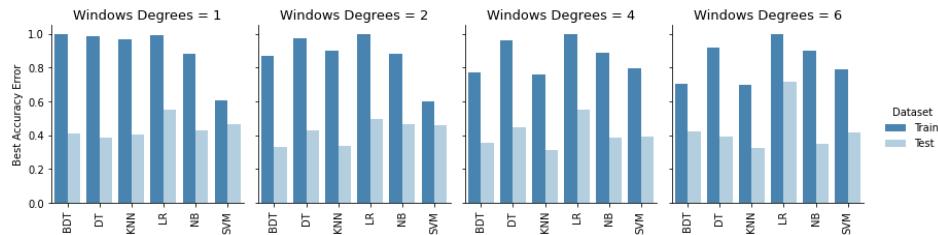


Figure 1. Classifier comparison of accuracy error by the window size.

It can be seen that LR outperforms the rest of the models, tuned with the hyperparameters, $C=1$, $penalty=none$, proving an accuracy error of 0.72. It is important to note also that LR shows a constant performance throughout all the cross-validation folds and hyperparameter combinations, and consequently presents a more reliable generalization. On the other hand, the KNN classifier shows a strong dependency on the train and test files selected, indicating that it is prone to overfitting.

Finally, the cross-validation confusion matrix for the LR model is presented in Table 1. The model performs better for the absence of defect due to the class imbalance.

Table 1. Cumulative cross-validation confusion matrix for quality prediction. Counts and percentage

Predicted	Good	4646	1649	Predicted	Good	81%	39%
	Bad	1110	2595		Bad	19%	61%
	Good	Bad			Good	Bad	
	Real				Real		

5. Conclusions

The model introduced focuses on predictive quality solutions for the FSW stations of the Ariane Pre-Final Assembly Line. The design proposed, developed, and benchmarked

builds upon a data processing and feature engineering layer. It establishes the spatial-time correlation, based on a moving window that determines the relation between process parameters timestamps and quality inspection position in degrees.

To predict the presence of the defect, a strategy based on model competition and benchmarking for different ML classifiers is presented. Employing two-step cross-validation, the performance of the different algorithms, hyperparameters and moving window sizes is evaluated. Larger moving windows obtain more balanced results between train and test data, therefore a compromise solution of 6-degree windows is selected. The presence of defects can be predicted with over 70% of accuracy with a LR model. There is an important decrease in the performance of test data compared to train data. Large differences may indicate that the model is still not completely able to generalize and perform under new welding conditions, indicating that models suffer from overfitting (what is learned in the training phase does not match new unseen data). This is due to two main reasons: limited available data together with different configurations of the FSW station in each dataset.

Overall, these results, although preliminary given the small dataset, show the potential of the predictive quality module carried out within the SESAME project for complex manufacturing systems in the aerospace industry which is particularly challenging due to the low production cadence and extreme quality requirements. Furthermore, it can provide very valuable insights to process experts to focus their attention on specific welding sections that need to be inspected for quality assurance.

6. Acknowledgment

This work has been carried out in the framework of the SESAME project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821875. The content of this paper reflects only the author's view; the EU Commission/Agency is not responsible for any use that may be made of the information it contains.

7. References

- [1] <https://sesame-space.eu>
- [2] Thomas WM, et al., Friction stir welding for the transportation industries, Mater. Des. 1997;18, 4–6.
- [3] Camps M., et al., Data-Driven Analysis of Friction Stir Welding for Aerospace Applications. In Artificial Intelligence Research and Development 2021 (pp. 181-184). IOS Press.
- [4] Welch P. The use of the fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoust. 1967; 15: 70-73.
- [5] Pearson K. On lines and planes of closest fit to systems of points in space. Philos. Mag. Lett. 1901. 2.11:559
- [6] Iglewicz B. Robust scale estimators and confidence intervals for location. Understanding robust and exploratory data analysis, 1983, 404: 431.
- [7] Boser BE, et al., A training algorithm for optimal margin classifiers. Proc. COLT. 1992: 144-152.
- [8] Fix E, et al., Discriminatory analysis. Nonparametric discrimination: Consistency properties. ISR, 1989, 57.3: 238-247.
- [9] Sutton CD. Classification and regression trees, bagging, and boosting. Handb. Stat, 2005, 24: 303-329.
- [10] Hosmer Jr. DW. Lemeshow, S.; Sturdivant, R. X. Applied logistic regression. John Wiley & Sons, 2013.
- [11] Murphy KP, et al. Naive bayes classifiers. University of British Columbia, 2006, 18.60: 1-8.
- [12] Hastie T, et al. The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

Efficiency and Reliability Enhancement of High Pressure Die Casting Process Through a Digital Twin

Pol TORRES^{a,1}, Albert ABIO^a, Raquel BUSQUÉ^b, Albert BRÍGIDO^b, Sylvia Andrea CRUZ^c, Manel DA SILVA^c and Francesc BONADA^a

^a*Eurecat, Centre Tecnològic de Catalunya, Unit of Applied Artificial Intelligence, Av. Universitat Autònoma, 23, 08290 Cerdanyola del Vallès, Spain*

^b*Eurecat, Centre Tecnològic de Catalunya, Product Innovation & Multiphysics Simulation Unit, Av. Universitat Autònoma, 23, 08290 Cerdanyola del Vallès, Spain*

^c*Eurecat, Centre Tecnològic de Catalunya, Metallic and Ceramic Materials Unit, Av. Universitat Autònoma, 23, 08290 Cerdanyola del Vallès, Spain*

Abstract. Several factors can contribute to the final part quality in a High Pressure Die Casting process, in terms of roughness, porosity and strength. The injection velocity, the cooling of the die and aluminium's inlet temperature are some of the factors that can have a higher effect on the part quality. The new advances on process digitalization, sensorization and simulation tools, combined with artificial intelligence techniques allow developing a functional Digital Twin aimed to monitor in near real time the evolution of the temperature and pressure during the production cycle to detect possible anomalies and predict the final part properties, reducing the required quality control.

Keywords. Digital Twin, Artificial Intelligence, Machine Learning, Numerical simulation.

1. Introduction

The concept of Digital Twin (DT) appeared for the first time in 2002 in the context of product lifecycle management [1]. A DT is defined as a digital mirror of a physical system created from real data which includes algorithms and decision making [2]. In this sense, a Digital Twin must involve three main features: real system, virtual modelling of the system (including visualization tools and algorithms), and bidirectional communication between the real system and the virtual model.

The use of solutions based on artificial intelligence (AI) is increasing due to the need to solve complex problems that are too difficult to address with conventional analytic tools. A DT, which can combine complex AI-based algorithms, together with analytic and visualization tools, can be applied in several environments. For instance, DTs are being applied to jet engines in the aeronautic sector to ensure an optimal predictive maintenance strategy and to evaluate the behaviour of the monitored asset in front of unexpected events and different climate conditions [3]. In the automotive sector a DT

¹ Pol Torres, Applied Artificial Intelligence Unit, EURECAT; E-mail: pol.torresalvarez@eurecat.org.

can be applied to crash tests and to improve autonomous driving [4]. In logistics and value chain control, a full production plant can be represented by a DT used to optimize the productivity [5], check the process quality, improve energy efficiency, and do demand prediction including several external variables such as natural disasters, pandemics, and political conflicts. More recently, the use of DT is being applied to control and improve the efficiency of electric batteries [6] and could be also very useful to control smart cities as they can analyse and predict the traffic state, the weather, and the electric demand on buildings through machine learning algorithms among other features [7].

This paper explains the ongoing work on the design and development of a Digital Twin for the High Pressure Die Casting process (HPDC) with AlSi₈Cu₃ performed in the Eurecat premises.

2. Methods

To set up the Digital Twin of the HPDC process it is required to obtain experimental data from the manufacturing process (machine, die, sensors) and send this data to a digital platform, where simulations-based and AI-based models can be applied, show visual information for supporting decision making, and return information/commands to the shopfloor. The data acquisition of the process is done by three sensors (two pressure sensors and one temperature sensor) placed in the inner side of the die where the part is casted. The plunger position and velocity, which control the injection process, are also registered. The sensors placed in the die and in the injection machine (Bühler Evolution 53D) send the information to an industrial PC via OPC, to be then ingested by the DT platform.

The near real time data visualization in this DT is done in 4D, showing the surface evolution of the temperature/pressure as function of time. To obtain a 4D representation of the process it is required to calculate previously a set of similar configurations by numerical simulation in InspireCAST. Considering the geometry of the part the injection process is simulated. Matching the real temperature and pressure from the sensors with the simulated ones, a data driven calibration can be established to generate all the simulated sample grid. The simulations provide relevant information of the injection process, such as temperature and pressure as function of time and the final porosity of the sample (part quality indicator). Despite the expensive computational time of these simulations, AI algorithms can be combined with simulations and experimental tests to provide accurate results in a very reduced time window.

To create the DT of the HPDC process, four AI-driven models are considered: anomaly detection, virtual sensors, quality prediction and process configuration prediction. The anomaly detection is performed by means of an unsupervised Isolation Forest model considering a 3% of outlier factor, which checks the goodness of each new cycle data. A Virtual Sensor (VS) is able to reproduce the trend of a real sensor (temperature or pressure in this case) through correlations with other parameters, and then reduces the need (and the cost) of using real sensors. Extra Trees Regressor [8], Random-Forest Regressor [9], and Support Vector Machine Regressor [10] algorithms with different combinations of hyperparameters have been tested. For quality and process configuration prediction multi-class classification algorithms are used. The aim of the quality prediction is to determine the quality of the sample, which is tagged by visual inspection of the expert operator from 1 to 4, being 4 the best quality. The process

configuration algorithm is developed to predict the configuration of the process (injection velocities v_1 and v_2 , and aluminum inlet temperature) from the sensor's information. Extra Trees Classifier, Random Forest Classifier and Gradient Boosting Classifier [11] algorithms have been tested in both cases. For all the models 80% of samples are used for training and 20% for test.

3. Results and discussion

The DT of the HPDC process is created with the AI-based algorithms showing a lower root mean squared error (RMSE) and higher accuracy score. Considering all the features extracted from the sensor's information, the Maximum Information Coefficient (MIC) is calculated, which expresses the correlation of the different parameters with the target. Table 1 shows the MIC for the sample quality and process configuration.

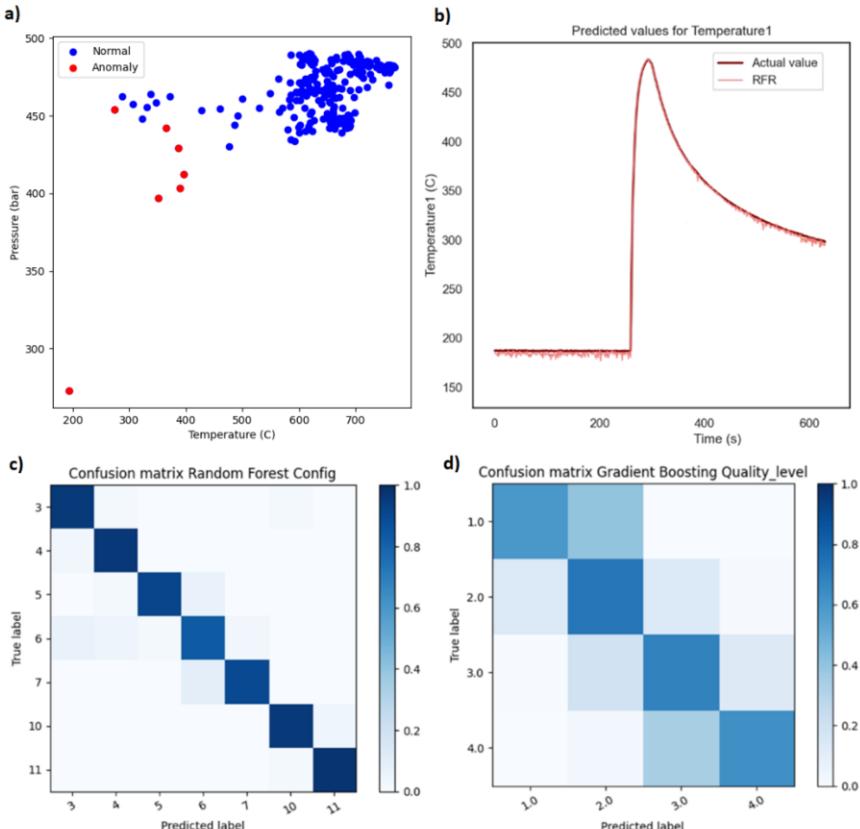


Figure 1. a) Anomaly detection prediction. b) Virtual sensor of Temperature. Confusion matrix for configuration prediction (c) and quality level (d).

For anomaly detection, the current developed algorithm has a score of 84%. As it can be observed in Figure 1a, there is a big cluster in the top right side that corresponds

to the parts without die refrigeration and with a quality of 3 and 4. The other spread samples correspond to samples with die refrigeration, which have a quality of 1 and 2. Although the samples are spared, a small cluster is observed and therefore the algorithm does not detect them as anomaly.

For the part temperature VS, the most relevant parameters are the pressure of sensor 1 ($MIC = 0.92$), the injection velocity ($MIC = 0.84$), and the plunger position ($MIC = 0.79$). Using these variables as input, the predictions with lowest RMSE (6°C) are found for the Random Forest Regressor. Figure 1b shows the prediction of the RFR compared with the real experimental data, where it can be observed that although there is some noise in the initial flat zone and during the cooling, the overall prediction reproduces the sensor's data.

Table 1. Higher correlations of the sample quality process configuration

Quality		Configuration	
Variable	MIC	Variable	MIC
Temperature increase	0.81	Initial Temperature	0.87
Max. Temperature	0.76	Max. Pressure 1	0.84
Initial Temperature	0.74	Pressure 1 increase	0.83
Cooling exponent	0.66	Max. Temperature	0.74

To predict the sample quality, the temperature increase experienced by the matrix is the most important parameter, while to predict the configuration of the sample the initial temperature, the maximum pressure and the pressure increase (at sensor 1) are the key factors. Based on Table 1, Figure 1c and Figure 1d show the results obtained for configuration and quality prediction, showing a score of 89.5% for the configuration (Figure 1c) and 68% for the quality (Figure 1d). Notice that although the score for quality prediction is lower, the error is limited to adjacent values (similar qualities).

4. Conclusions

In this work the preliminary results of the ongoing DT development of the HPCS process implemented at Eurecat's Laboratory facilities are presented. The methodology developed in this project is done in a way that can be extended to other industrial processes with minor modifications of the pipeline. The results show that the AI-based models can predict with high accuracy some process features like the produced part quality and estimate the process configuration parameters from few sensor's information. In addition, a temperature virtual sensor can be obtained from training regressors with the information of pressure of the die, injection velocity and plunger position. In the same way, the pressure of the die can be obtained from the temperature and the other two parameters.

The next steps will focus on improving of the prediction models to enable predictive maintenance solutions, and to extend the testing of the algorithms with different sample geometries and other aluminium alloys.

Acknowledgements

This work was financially supported by the Catalan Government through the funding grant ACCIÓ-Eurecat (Project PRIV DIGITS)". Albert Abio acknowledges the financial support of Eurecat's "Vicente López" PhD grant program.

References

- [1] Kritzinger W, Karner M, Traar G, Henjes J, Sihn W, Digital Twin in manufacturing: A categorical literature review and classification, IFAC-Papers OnLine. 2018. 51 (11): 1016-1022.
- [2] Rosen R, Wichert G, Lo G, Bettenhausen K. About The Importance of Autonomy and Digital Twins for the Future of Manufacturing. In IFAC-Papers OnLine. 2015. 48 (3): 567–572.
- [3] Tuegel E, Ingraffea A, Eason T, Spottswood S. Reengineering Aircraft Structural Life Prediction Using a Digital Twin. International Journal of Aerospace Engineering. 2011. 2011: 1687-5966
- [4] Veledar O, Damjanovic-Behrendt V, Macher G. Digital Twins for Dependability Improvement of Autonomous Driving. [Communications in Computer and Information Science](#) book series (CCIS, volume 1060): Springer; 2019. p. 415–426
- [5] Zhang H, Liu Q, Chen X, Zhang D, Leng J. A Digital Twin-Based Approach for Designing and Multi-Objective Optimization of Hollow Glass Production Line. IEEE Access. 2017. 5. 26901-26911.
- [6] Wu B, Widanage W. D, Yang S, Liu X. Battery digital twins: Perspectives on the fusion of models, data and artificial intelligence for smart battery management systems. Energy and AI. 2020. 1. 100016.
- [7] Deren L, Wenbo Y, Zhenfeng S. Smart city based on digital twins. Computational Urban Science. 2021. 1(4). 2730-6852.
- [8] Geurts P, Ernst D and Wehenkel L. Extremely randomized trees. Mach. Learn. 2006. 63. 3-42.
- [9] Breiman L. Random Forests. Mach. Learn. 2001. 45. 5-31
- [10] Smola A, Schölkopf B. A tutorial on Support Vector Regression. Statistics and Computing archive. 2004. 14 (3). 199-222.
- [11] Friedman J. H.. Stochastic Gradient Boosting. Technical Report, Stanford University, Stanford, 1999.

An Agent-Based Simulation Framework for Firefighters Training

Jordi SABATER-MIR ^{a,1}, Ignasi CAMPS-ORTIN ^a and Cristian COZAR-ALIER ^a

^a*IIIA, CSIC, Campus UAB, E-08193. Bellaterra, Catalonia (Spain)*

Abstract. In this paper, we introduce RHYMAS, a simulation framework that uses a multiagent approach and allows the creation of simulated emergency scenarios that mix artificial agents and humans. Those humans interact with the simulation using different types of immersive technologies, including virtual reality, simtables or virtual cave environments. The RHYMAS framework is being designed specifically for training firefighters and is being developed in collaboration with the “Escola de bombers i protecció civil de Catalunya”.

Keywords. computer-based simulation, multiagent systems, autonomous agents

1. Introduction

The use of computer simulations as a tool for training qualified professionals on the skills needed in their jobs has a long tradition in many areas (healthcare, military, emergencies, transport, etc.). In general, these are areas where reproducing the training scenario in the real world is costly, dangerous or directly unfeasible. Recently, the interest in these kinds of simulations has greatly increased due to the improvement and cost reduction of immersive technologies like Virtual Reality (VR) and Augmented Reality (AR) that reduce the distance between reality and the simulation.

The RHYMAS framework is a framework under development in collaboration with the “Escola de bombers i protecció civil de Catalunya” specifically designed for firefighters training. The framework is being designed to allow for large training simulations that combine different tactical and strategic levels and that at the same time are easy to enact, easy to control and with a high training capacity. This is achieved approaching the simulations using a multi-scale paradigm and adding autonomous agents and multiagent systems technology to reduce the amount of human (simulation operators) intervention during the training sessions.

2. Example scenario

In this section we will present an example of the kind of complex training scenarios that RHYMAS is targeting. In this example, the emergency situation (based on a real case)

¹Corresponding Author: Jordi Sabater-Mir, IIIA, CSIC, Campus UAB, E-08193. Bellaterra, Catalonia (Spain); E-mail: jsabater@iiia.csic.es.

starts at the Barcelona's harbour where a passenger ship hits a harbour crane that falls on a group of containers and starts a fire. The ship's crew manoeuvres quickly and takes the ferry to a safer area as far away from the scene as possible but with many injured among the passengers. In the meantime, a smoke column comes out of the fire in the containers, creating a toxic cloud that is spreading over the city.

At this point is where the training scenario starts. Initially there are two sectors: in sector 1 there is the ship with 120 injured people and in sector 2 there is the fire in the containers with the smoke column. A scenario like this has an enormous complexity. Figure 1 shows a recreation of the different human actors that could participate in the simulation.



Figure 1. Recreation of a human-in-the-loop complex emergency simulation.

At the first area (number 1 in Figure 1) you have groups of firefighters that are dealing face-to-face with the emergency (injured people, fire, etc.). At this level, the simulations are localised in a small area but need a lot of detail (convincing avatars and fire recreation, high graphical detail, etc.). Those are the firefighters that in the example would be dealing with the injured people and the ship in sector 1 and would be in front of the fire in sector 2.

The second area is populated by officers that direct groups of firefighters (number 2 in Figure 1). The groups of firefighters these officers are leading are humans participating in the simulation in the first area or groups of avatars controlled by the simulation engine. In the example scenario there would be at least one officer for each sector but if things go wrong, and the number of firefighters has to increase, a new officer layer is added to coordinate the different groups. The perception of the emergency scenario that the officers have at this level is from a certain distance but they are still located at the place of the emergency. Simulation detail needs to be quite high but not so detailed as it is in the first area.

In the example scenario there is also the added problem of a toxic smoke column that, after a while, spreads dangerously over the city. This requires the creation of a third level, a new command centre that will deal with that aspect of the emergency (number 3 in Figure 1). In this case, the perspective those officers have of the emergency is at

a map level. By looking at the evolution of the smoke, the officers at this level provide instructions to the officers at level 2.

Finally, the last simulation level (number 4 in Figure 1) is composed by high rank officers. They receive reports from the other levels and determine the main course of action.

3. The RHYMAS architecture

Figure 2 shows the main elements of the RHYMAS architecture. The whole simulation is composed by a set of federated simulations. Each one of these simulations has its own independent simulation engine that is able to exchange information about its state with the other simulations. The different simulations are supervised by a sim operator that is responsible of high level management tasks.

A federated simulation is, at the same time, composed by the backend and one or several frontends. The backend implements the simulation logic and is responsible of communicating with/receiving information from the other simulations Backends. The frontends have two main functions: they provide different user views on that simulation (showing to the user the elements of the simulation that are relevant for him/her) and they can function as solvers for some specific aspects of the simulation, specifically physics, movement and perception. Separating this two aspects of the simulation has several advantages like the possibility of using state of the art game engines for the frontends but without conditioning the logic of the simulation to the game engine's idiosyncrasy.

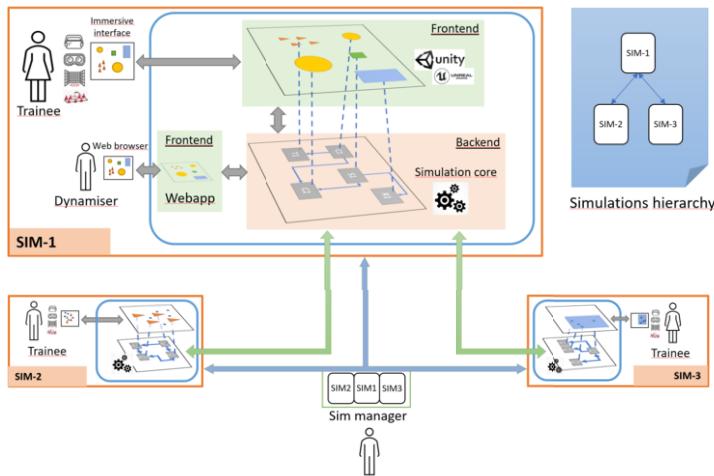


Figure 2. General RHYMAS architecture.

The counterpart of this separation between the Backend and the Frontends is that it increases the complexity of the architecture. An entity that is part of a simulation is split in two parts, the one in the Backend that we can see as the 'brain' of the entity and the one in the Frontend that we can see as the 'body' (sensors and actuators). The two parts of the same entity communicate between them using messages. So for example, when the Frontend part detects a collision with another object, if this is relevant for the

simulation, it will send a message to the Backend part notifying it, probably indicating the ID of the other entity and other relevant information regarding that collision. On the other hand, if the Backend part decides that the entity has to move to a specific location, it will send a message to the Frontend part with the target location. The Frontend part will start the movement in the 3D environment taking care of the details, like for example which animation needs to be played or the calculation of the pathfinding. The Backend will be notified only if there is something associated to the movement that is relevant to the logic of the simulation.

A RHYMAS Backend is implemented as a directed graph where the nodes are the agents (elements) of the simulation (fires, firefighters, firetrucks, victims, etc.) and the edges the relations that exist among them. These relations represent any kind of influence that an agent exerts over another agent at a specific time step. A relation stores all the details that are relevant about that influence. Agents and relations are created and removed on demand following the evolution of the simulation. The Backend simulator engine is implemented as a traditional synchronous simulation with a main loop that each step asks sequentially the agents and relations to update their internal state. The state of an agent is the result of its internal model, the state of the incoming relations and the messages received from its counterparts in the Frontends. As a consequence of this update, apart from changing its internal state, the agent can create new relations and modify the details of the outgoing existing relations. Each simulation step the process is repeated.

4. Current state and future work

The framework is still under development. At the moment of writing these lines, we are working on a first prototype that implements the ideas described in this paper in a single simulation scenario like the ones used in the “Escola de bombers de Catalunya”. The prototype is being implemented using Python for the Backend and the Unity game engine and different web frameworks for the Frontends.

The next step will be extending the framework to allow scenarios with multiple simulations at the same scale level. This will imply solving the problem of the communication between simulations and their synchronisation. Finally, we will extend the functionalities of the framework to include scenarios with multiple simulations at different scales like the one described in section 2.

Acknowledgements

This research has been funded by the Ministerio de ciencia e innovación “Programa Estatal de I+D+i Orientada a los Retos de la Sociedad” through the project “RHYMAS-Real-time Hybrid Multiscale Agent-based Simulations for emergency training” (PID2020-113594RB-100). Ignasi Camps is financed by a IIIA-CSIC JAE Intro Scholarship. We thank the people at the “Escola de bombers i protecció civil de Catalunya” for their advice in the scenarios definition.

Data Driven Predictive Models Based on Artificial Intelligence to Anticipate the Presence of *Plasmopara viticola* and *Uncinula necator* in Southern European Winegrowing Regions

Marta OTERO^{a,1}, Luisa Fernanda VELASQUEZ^a, Boris BASILE^b, Jordi Ricard
ONRUBIA^a, Alex Josep PUJOL^a, Josep PIJUAN^a

^a Eurecat, Centre Tecnològic de Catalunya, Unit of Applied Artificial Intelligence,
Science and Technology Park of Lleida, Building H3, 25003 Lleida, Spain

^b Department of Agricultural Sciences, University of Naples Federico II, 80055 Naples,
Italy

Abstract. Downy and powdery mildews are two of the main diseases threatening grapevine cultivation worldwide caused by the phytopathogens *Plasmopara viticola* and *Uncinula necator*, respectively. These diseases may cause severe damage to grapevines by inducing wilting of plant organs, including bunches, especially when vines are untreated. This fact, together with the widespread of these pathogens due to the large extensions of land dedicated to grapevine monoculture, makes necessary to develop new predictive modeling tools that allow anticipating disease appearance in the vineyard, minimizing the losses in fruit yield and quality, and helping farmers in defining appropriate and more sustainable disease management strategies (fungicides applied at the right time and dose). For this purpose, farms located in three countries (Portugal, Spain, and Italy) were selected to study the relationship between the microclimatic characteristics of the plots, the phenological stage of the plants throughout the annual cycle, and the presence of both pathogens using different Machine and Deep Learning classification algorithms: Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, K-Nearest Neighbors, Naïve Bayes, Support Vector Machines, and Deep Neural Networks. The results showed that, after an entire annual grapevine cycle, the best performing models were Support Vector Machines for downy mildew and Random Forest for powdery mildew, providing a prediction accuracy of more than 90% for the infection risk and more than 80% for the treatment recommendation. These models will be fine-tuned during two additional vegetative seasons to ensure their robustness and will receive short- and medium-term climatological and phenological forecasts to make recommendations. The preliminary results obtained show that these models are a promising tool in the field of plant disease prevention and resource saving.

Keywords. disease anticipation, resource saving, fungicide management, prevent economic losses, Machine learning.

1. Introduction

Europe has always had a strong wine culture for thousands of years, making many of its viticultural regions world leaders in wine or fresh fruit production. Although new countries have been joining vine production for self-consumption and export, the

¹ Corresponding author: Marta Otero, Unit of Applied Artificial Intelligence, Eurecat, Centre Tecnològic de Catalunya, Science and Technology Park of Lleida, Building H3, Lleida, Spain. E-mail: marta.oter@eurecat.org

European Union continues to be the pioneer in world production, reaching figures of more than 60% [1]. France, Italy, and Spain are among the countries with the highest wine production in the European Union. Grapevines (*Vitis vinifera L.*) are usually located in slopy areas or on flat terrains and typically grow under Mediterranean climatic conditions. Due to the characteristics of the farms, monoculture is usually encouraged, and often a single variety is grown. This cultivation model favors the appearance of diseases and can cause catastrophic consequences in terms of production [2]. Two of the main diseases affecting grapevine production are downy and powdery mildews caused by *Plasmopara viticola* and *Uncinula necator*, respectively. In recent years, several authors have tried to approach plant disease detection with Artificial Intelligence [3,4] obtaining promising results from different perspectives. However, these studies were limited to very specific climatic zones and cultivars without taking into account neither the evolution of the disease in different grapevine varieties and climatic conditions nor the interaction between fungal development and grapevine phenology. Therefore, the objective of the present work was to monitor the epidemiology and management of *P. viticola* and *U. necator* in different growing areas and grapevine varieties to understand and predict the behavior of both phytopathogens. This will allow anticipating their appearance and designing an efficient management protocol of the vineyards, reducing the use of resources such as water and fungicides by applying the treatments at an appropriate dose and date and preventing economic losses.

2. Materials and Methods

2.1. Selection of sampling plots, data collection and preprocessing

The study areas are centered in the Mediterranean and Atlantic region of vine production in the European Union: Quinta do Ataíde (Portugal), L'Aranyó (Spain), and Mirabella Eclano (Italy). In each of the experimental plots the following disease management treatments were compared: i) an unsprayed control, ii) a treatment where to apply the recommendations generated by the models and iii) a treatment where diseases were managed conventionally. Each treatment consisted of a total of 20 plants divided into 4 blocks. Disease monitoring during the first growing season was conducted during the months of March-October 2021, making field visits every 7 days. On each measuring date, a total of 100 leaves per treatment were visually inspected to measure the percentage of affected leaves and the percentage of affected leaf blade. With respect to the treatments applied in the field to control the diseases, the historical data of product applications and doses supplied during 2021 was collected. Depending on the vine phenological stage, an adjustment was made to the predictions in terms of the volume of mixture applied in the field based on previous knowledge, optimizing the use of water used to apply treatments.

The independent features used in the model were climatology and the evolution of phenology throughout the year. The former comprises daily data obtained from *in situ* sensors that measure the microclimatic plot conditions, while the latter is based on in-field crop observations by the farmers. Climatological differences between plots were determined by Repeated Measures Analysis of Variance using the Least Significance Difference as a post hoc test. Finally, different feature engineering techniques were used to obtain the best possible data quality. In addition, quantitative disease occurrence data were translated into qualitative data to provide low-, medium- and high-risk alerts and propose treatment dosages.

2.2. Initial model selection and effectiveness evaluation

Once the data were ready to be introduced into a model, eight different classification Machine and Deep Learning models (Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, K-Nearest Neighbors, Naïve Bayes, Support Vector Machines and Deep Neural Networks) were studied using the scikit-learn library [5] in Python3. For this purpose, a battery of different hyperparameters was prepared for each of the models, and class compensation by stratification was performed during the separation of the train and test datasets. All models were trained with a 10-fold cross validation and, when necessary, the features were standardized. The degree of success in predicting the models was expressed in terms of accuracy.

3. Results and discussion

Regarding the characterization of the study regions, Table 1 shows the average relative air humidity, air temperature, and precipitation for the period 2017-2020 and only for the year 2021. The results indicate that (a) the average air temperature was significantly different in the three regions, (b) L'Aranyó is also the site with less rainfall, and (c) in 2021 there was a slight increase in relative air humidity with respect to the previous four years, which *a priori* could favor the appearance of phytopathogens.

Table 1. Characterization of the most important environmental features for the growth of phytopathogens in the three study regions from 2017 to 2020 (top) and in 2021 (bottom): daily mean humidity, mean temperature, and precipitation. The values represented correspond to the mean of the values and their standard deviation. The asterisk indicates significant differences between the corresponding regions for each feature.

Period	Region	Air humidity (%)	Air temperature (°C)	Precipitation (mm)
2017 - 2020	L'Aranyó	69.643 ± 14.220*	15.288 ± 7.492*	1.182 ± 5.925*
	Mirabella Eclano	70.055 ± 14.671*	14.910 ± 6.893*	2.424 ± 6.618
	Quinta do Ataíde	66.908 ± 15.759*	16.729 ± 6.788*	1.433 ± 4.496
2021	L'Aranyó	71.145 ± 14.408	14.792 ± 7.442*	0.751 ± 2.998*
	Mirabella Eclano	72.568 ± 24.620	18.009 ± 6.908*	2.520 ± 6.315
	Quinta do Ataíde	67.630 ± 13.762*	16.185 ± 6.348*	1.288 ± 3.950

Considering the predictive reliability offered by the models, it is possible to observe that, in general terms, a prediction accuracy of over 90% was achieved for the different levels of disease and over 80% for the treatment recommendation (Table 2). To select the most appropriate algorithm, disease classification models were prioritized, seeking not only high degrees of accuracy, but also a minimum difference between the train and test sets. Therefore, the Support Vector Machines (SVM) model for downy mildew and the Random Forest (RF) model for powdery mildew were finally selected. The coefficients of the feature importance shown by these models agreed with those described in the literature, since the SVM model showed the minimum/average air temperature and the degree of humidity as determining features for classification, which is consistent with the oospores being completely dependent on environmental and foliar humidity for their penetration into the plant [6,7]. In addition, the RF model showed that the most important feature for classification was air temperature, which is also in agreement with all the literature studied, in which temperature is established as a critical feature for secondary infection [8,9]. Both models, but especially the SVM, emphasized the importance of the phenological stage of the plant when classifying the risk of infection.

Table 2. Average of the accuracy values of the train-validation/test sets for each of the models developed. A_d means accuracy of the disease model and A_t means accuracy of the treatment model.

Model	Downy mildew A_d train/test	Downy mildew A_t train/test	Powdery mildew A_d train/test	Powdery mildew A_t train/test
Logistic regression	0.970 / 0.964	0.874 / 0.856	0.978 / 0.956	0.894 / 0.837
Decision tree	0.973 / 0.950	0.859 / 0.878	0.976 / 0.926	0.898 / 0.875
Random forest	0.977 / 0.957	0.870 / 0.878	0.978 / 0.963	0.911 / 0.881
Gradient Boosting	0.975 / 0.965	0.875 / 0.863	0.981 / 0.941	0.913 / 0.889
K-nearest neighbors	0.965 / 0.958	0.868 / 0.870	0.976 / 0.949	0.885 / 0.881
Support Vector Machines	0.965 / 0.964	0.882 / 0.863	0.978 / 0.956	0.889 / 0.859
Naïve Bayes	0.934 / 0.901	0.816 / 0.814	0.976 / 0.933	0.819 / 0.815
Deep Neural Network	0.965 / 0.944	0.868 / 0.798	0.969 / 0.926	0.882 / 0.837

However, despite the promising results, it is advisable to continue improving the models in the following years with the incorporation of more data from subsequent seasons that provide greater variance to the sample, as well as data from regions with different environmental conditions (e.g., regions with cold climates) to those already present for model validation. Finally, thanks to the predictions obtained from mature climatic and phenological models, which will serve as new inputs to the models, it will be possible to make short- and medium-term forecasts and treatment recommendations for growers.

Acknowledgements

We would especially like to thank Familia Torres Wines, Mastroberardino Società Agricola Srl, and Symington Family Estates for providing the vineyards for these experiments. We would also like to thank Mr. Federico Oldani from LINKS Foundation for facilitating access to the historical data in an automated way. L.F. Velasquez is a fellow of Eurecat's "Vicente López" PhD grant program. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 869565.



References

- [1] Pink M. The sustainable wine market in Europe - introduction to a market trend and its issues. *Acta Scientiarum Polonorum Oeconomia*. 2015;14(2).
- [2] Shipton PJ. Monoculture and Soilborne Plant Pathogens. *Annu. Rev. Phytopathol.* 1977;15(1): 387-407.
- [3] Shruthi U, Nagaveni V, Raghavendra BK. A Review on Machine Learning Classification Techniques for Plant Disease Detection. 5th International Conference on Advanced Computing & Communication Systems (ICACCS); 2019 March 15-16; Coimbatore, India: IEEE; c2019. p. 281-84.
- [4] Chen M, Brun F, Raynal M, Makowski D. Forecasting severe grape downy mildew attacks using machine learning. *PLOS ONE*. 2020;15(3): e0230254.
- [5] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011;12: 2825-30.
- [6] Lalancette N, Madden LV, Ellis MA. A quantitative model for describing the sporulation of *Plasmopara viticola* on grape leaves. *Phytopathology*. 1988;78:1316-21.
- [7] Orlandini S, Massetti L, Dalla Marta A. An agrometeorological approach for the simulation of *Plasmopara viticola*. *Comput. Electron. Agric.* 2008;64:149-61.
- [8] Chellemi DO, Marois JJ. Development of a demographic growth model for *Uncinula necator* by using a microcomputers spreadsheet program. *Phytopathology*. 1991;81:250-54.
- [9] Lu W, Newlands NK, Carisse O, Atkinson DE, Cannon AJ. Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards. *Agronomy*. 2020;10:622-50.

Towards and Efficient Algorithm for Computing the Reduced Mutual Information

Martí RENEDO-MIRAMBELL^a and Argimiro ARRATIA^{a,1}

^a*Soft Computing Research Group (SOCO)*

at Intelligent Data Science and Artificial Intelligence Research Center

Department of Computer Sciences,

Polytechnical University of Catalonia, Barcelona, Spain.

marti.renedo@gmail.com, argimiro@cs.upc.edu

Abstract. In [1], Newman et al. introduced the Reduced Mutual Information (RMI), a measure of the similarity between two partitions of a set useful in clustering and community detection. The computation of RMI requires counting the amount of contingency tables with fixed row and column sums, a #P-complete problem, for which the authors suggest to use analytical approximations that work in general, but for other not so pathological cases these give highly inaccurate approximations. We propose a hybrid scheme based on combining existing Markov chain Monte Carlo methods with analytical approximations to make more accurate estimates of the number of contingency tables in all cases.

Keywords. mutual information, contingency tables, clustering, Markov chain Monte Carlo

1. Introduction

The computation or approximation of the number of contingency tables with fixed row and column sums is a necessary step for the computation of Reduced Mutual Information (RMI). This is a #P-complete problem (see e.g. [2]), so we don't have any algorithm to perform the exact computation efficiently, which rules it out for even moderately sized networks. When introducing the RMI [1], Newman et al. suggest using analytical approximations, but they have important limitations. Particularly, they don't give accurate results when row and column sums contain numerous small elements (instead, the approximation is accurate when the contingency tables are very dense). On the other hand, it is possible to use a Markov chain Monte Carlo method, as described in section 2 , but it is much slower to compute. The idea behind our approach is to separate the part of the table for which the analytical formula is accurate, and use that to then obtain the result with fewer steps of the Monte Carlo method.

Reduced Mutual Information. Given r and s two labelings of a set of n elements, the Reduced Mutual Information is defined as:

¹Corresponding Author: Argimiro Arratia, e-mail: argimiro@cs.upc.edu

$$\text{RMI}(r; s) = I(r; s) - \frac{1}{n} \log \Omega(a, b). \quad (1)$$

where $\Omega(a, b)$ is a integer equal to the number $R \times S$ of non-negative integer matrices with row sums $a = \{a_r\}$ and column sums $b = \{b_s\}$ (i.e., contingency tables). In practice, computing or at least approximating $\Omega(a, b)$ with enough accuracy is the main challenge in obtaining the Reduced Mutual Information of two partitions.

2. Analytical approximation

The following approximation works in cases where the numbers of clusters R and S are relatively small relative to the total number of elements, resulting in very populated clusters. Let a and b vectors of lengths R and S respectively be the margins of the contingency table, and $\Omega(a, b)$ the corresponding number of contingency tables. Also, define:

$$w = \frac{n}{n + \frac{1}{2}RS}, \quad x_r = \frac{1-w}{R} + \frac{wa_r}{n}, \quad y_s = \frac{1-w}{S} + \frac{wb_s}{n}, \quad \mu = \frac{R+1}{R\sum_s y_s^2} - \frac{1}{R}, \text{ and}$$

$$v = \frac{S+1}{S\sum_r x_r^2} - \frac{1}{S}. \text{ Then:}$$

$$\begin{aligned} \log \Omega(a, b) \simeq & (R-1)(S-1) \log(n + \frac{1}{2}RS) + \frac{1}{2}(R+v-2) \sum_s \log y_s \\ & + \frac{1}{2}(S+\mu-2) \sum_r \log x_r + \frac{1}{2} \log \frac{\Gamma(\mu R)\Gamma(v S)}{[\Gamma(v)\Gamma(R)]^S [\Gamma(\mu)\Gamma(S)]^R}. \end{aligned} \quad (2)$$

However, this approximation can become highly inaccurate when the conditions aren't met. This can easily happen, for example, when a relatively high number of vertices are left isolated forming their own clusters, even if the rest of the clusters are large.

3. Monte Carlo approximation

An alternative approach is to use a Monte Carlo method to estimate the number of contingency tables by successively iterating over the set of solutions using an appropriately defined Markov chain. The method, introduced in [3], uses a nested chain of subsets $\Sigma_{ab} = H_1 \supset H_2 \supset \dots \supset H_t$. Then, Monte Carlo sampling is used to estimate each ratio $|H_i|/|H_{i+1}|$, which will allow the estimation of the whole set by just being able to enumerate H_t , which will be small (more specifically, it will contain a single element).

Random walk. First let's define a random walk on the set Σ_{ab} of matrices with row sums a and column sums' b . Let $M \in \Sigma_{ab}$. A pair of rows i_1, i_2 and columns j_1, j_2 is selected randomly. Then, $M' \in \Sigma_{ab}$ is obtained by adding 1 to the $(i_1, j_1), (i_2, j_2)$ elements and subtracting 1 to the $(i_1, j_2), (i_2, j_1)$ elements, or viceversa, each of the two possibilities with probability $\frac{1}{2}$. This gives a connected, symmetric, aperiodic Markov chain on Σ_{ab} .

Subset chain. Let $M \in \Sigma_{ab}$. Then, define $[\Sigma_{ab}|M; (k, l)]$ the subset of Σ_{ab} containing only tables that match M in all positions strictly preceding (k, l) in the lexicographic order. Then, if (k', l') succeeds (k, l) , then $[\Sigma_{ab}|M; (k', l')] \subseteq [\Sigma_{ab}|M; (k, l)]$. This gives

a chain of subsets $\Sigma_{ab} = [\Sigma_{ab}|M; (1, 1)] \subseteq \dots \subseteq [\Sigma_{ab}|M; (r, s)]$. The following result is proved in [3].

Theorem 3.1 *The random walk on $[\Sigma_{ab}|M; (k, l)]$ is ergodic and has uniform stationary distribution for all $M \in \Sigma_{ab}$.*

4. Hybrid analytical Monte Carlo approximation

We redefine the subset chain of the Markov Monte Carlo method to reduce its length by estimating the size of the biggest subset we can have analytically. We want to concentrate all the denser communities on one corner of the matrix, so a and b are sorted in ascending order. Then, divide the matrix into four blocks Q_1, Q_2, Q_3, Q_4 such that $|(\Sigma_{Q_4})_{ab}|$ can be estimated analytically. Of course, it is not possible to extend this estimation directly using the method described in section 2 because not all elements of Q_1, Q_2 and Q_3 precede those of Q_4 unless Q_4 has only one row.

Order relation. Here we will define an order in which to traverse the matrix M of $R \times S$ elements, or equivalently, a total order relation on the set $[R] \times [S]$. Let \prec , and \preceq denote the lexicographical order (the strict and non-strict versions respectively), and $p \in [R] \times [S]$ the element at the lower right corner of Q_1 . Then, we define the strict order relation \sqsubset as follows:

$$\begin{aligned} x \sqsubset y &\iff x \prec y && \text{if } x, y \in Q_1 \cup Q_2 \\ x \sqsubset y &&& \text{if } x \in (Q_1 \cup Q_2), y \in (Q_3 \cup Q_4) \\ (x_1, x_2) \sqsubset (y_1, y_2) &\iff x_2 < y_2 \text{ or } (x_2 = y_2 \text{ and } x_1 < y_1) && \text{if } x_1, y_1 > p_1 \end{aligned}$$

In other words, \sqsubset puts the elements of Q_1 and Q_2 first in lexicographical order, and then those of Q_3 and Q_4 in a variation of the lexicographical order that goes from left to right and top to bottom in that order. That puts all elements of Q_4 after any element of Q_1, Q_2 , and Q_3 . We will denote \sqsubseteq the non-strict version of the strict order relation \sqsubset .

Hybrid algorithm. With the order relation \sqsubseteq , we can define $[\Sigma_{ab}|M; (k, l)]_\sqsubseteq$ as the subset of Σ_{ab} containing tables that match M in all positions strictly preceding (k, l) in the \sqsubseteq order. To obtain a random walk on $[\Sigma_{ab}|M; (k, l)]_\sqsubseteq$, we just need to uniformly select a pair of rows $i_1 < i_2 \leq R$ and columns $j_1 < j_2 \leq S$ such that $(k, l) \sqsubseteq (i_1, j_1)$. Only elements that succeed (k, l) in the \sqsubseteq order will be modified by the random walk. We have

Corollary 4.1 *of theorem 3.1.* *The random walk on $[\Sigma_{ab}|M; (k, l)]_\sqsubseteq$ is ergodic and has uniform stationary distribution for all $M \in \Sigma_{ab}$.* \square

Then, the resulting algorithm can be described as follows:

- Rearrange the rows and columns of M so that their sums are in ascending order.
- Determine $p = (p_1, p_2)$ the position of the upper left corner of Q_4 . This is the cutoff point between the small and large communities, here we are using the first row and column with size > 1 .
- Estimate the values $|H_1|/|H_2|, |H_2|/|H_3|, \dots, |H_{q-1}|/|H_q|$, where $H_q = [\Sigma_{ab}|M; (p_1, p_2)]_\sqsubseteq$, with the Markov chain Monte Carlo method.
- Approximate H_q with the analytical formula described in section 2.
- Multiply the chain of fractions from the previous steps to obtain $H_1 = \Sigma_{ab}$.

5. Experiments and discussion

To test the standard Markov chain Monte Carlo and the hybrid algorithms, we use two vectors to set the margins of the tables, and execute both. The chosen vectors are: $a = (20, 10, 10, 5, 5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ and $b = (10, 10, 5, 5, 5, 5, 5, 1, 1, 1, 1, 1, 1, 1, 1)$, which correspond to a network of 50 nodes. Even in such a relatively small network, exact counting algorithms are not practical. Using the analytical approximation alone, the results are not meaningful because of the presence of a few isolated vertices, which makes the contingency tables less dense.

For the hybrid method, the matrix is split such that Q_4 sub-matrix is formed by the rows and columns with sum greater than 1. The computation took 24.2 seconds, almost twice as fast as the standard Markov chain Monte Carlo method (41.32 seconds). The term $\frac{1}{n} \log \Omega(a, b)$ estimated with each method differs by less than 0.01, so there is not a significant loss of accuracy when the method is used for the computation of the Reduced Mutual Information. In comparison, using only the analytical formula on the whole matrix produces an estimation that is off by over 0.3, which is clearly too inaccurate to obtain any meaningful estimation of the Reduced mutual Information.

If we instead study a case with fewer single element labels: $a = (25, 25, 15, 10, 4, 1)$ and $b = (25, 20, 15, 9, 8, 8, 1, 1, 1)$, the difference is much more apparent with the hybrid method taking 2.98 seconds compared to 37.41 of the standard Monte Carlo.

It is worth noting that the implementation of the Markov chain uses a naive sampling method that doesn't take advantage of the sparsity of the matrix in some areas. When the chosen elements that have to be decreased by one are already 0, the matrix remains invariant for that step of the chain. Then, when the matrix is very sparse and most of the steps are going to be invariant, it is possible to optimize the process by simply simulating the number of invariant steps before the matrix changes with a geometric distribution, and then sampling only from the rows and columns which will result in a step that modifies the matrix. This optimization would be a lot more beneficial on the sparser parts of the matrix (Q_1, Q_2, Q_3) and much less on the Q_4 sub-matrix, which would benefit the hybrid method more than the standard Monte Carlo method.

The implementation of the RMI measure presented here will be released as part of the `clustAnalytics` R package [4], with the goal to provide a readily available tool for cluster analysis on networks.

References

- [1] Newman MEJ, Cantwell GT, Young JG. Improved mutual information measure for clustering, classification, and community detection. *Phys Rev E*. 2020 Apr;101:042304. Available from: <https://link.aps.org/doi/10.1103/PhysRevE.101.042304>.
- [2] Dyer M, Kannan R, Mount J. Sampling contingency tables. *Random Structures & Algorithms*. 1997;10(4):487-506.
- [3] Diaconis P, Gangolli A. Rectangular Arrays with Fixed Margins. In: Aldous D, Diaconis P, Spencer J, Steele JM, editors. *Discrete Probability and Algorithms*. Springer New York; 1995. p. 15-41.
- [4] Renedo-Mirambell M. `clustAnalytics`: Cluster Evaluation on Graphs; 2022. R package version 0.3.1. Available from: <https://CRAN.R-project.org/package=clustAnalytics>.

Bootstrap-CURE Clustering: An Investigation of Impact of Shrinking on Clustering Performance

Ashutosh KARNA^{a,1}, Karina GIBERT^a

^aKnowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract. Hierarchical clustering is one of the most popular techniques in unsupervised segmentation. However, since it has quadratic complexity as it is based on pairwise distance matrix construction, it tends to be less used with really large data cases. *CURE* clustering tackles this challenge by accelerating the process through a first hierarchical clustering over a smaller sample from which a set of representative points of resulting clusters is obtained and used to estimate the cluster shape. A *KNN* process with those representative points allows completing the cluster assignment to the remaining points. This clustering technique scales the hierarchical clustering to large datasets. This work is in continuation of the earlier research, *Bootstrap-CURE* which uses repeated samples in the first part of the process and gains both robustness and representativeness. Also, the proposed approach uses a criterion for automatic identification of the number of clusters from a dendrogram, so that the bootstrap samples can be automatically processed. In this paper, the concept of shrinkage is proposed as a hyperparameter to the *Bootstrap-CURE* clustering approach. The inclusion of shrinkage brings the proposed clustering technique closer to the original *CURE* clustering. The impact of shrinkage on the overall performance of *Bootstrap-CURE* is further explored. A real-life use case from 3D printers is presented to illustrate the performance of the proposed clustering.

Keywords. CURE clustering, Hierarchical clustering, Cluster validity indices, Bootstrapping, Dendrogram

1. Introduction

Clustering is one of the most important machine learning techniques to discover the hidden patterns in a dataset. The quality of clustering results depends on both the similarity metric and the implementation technique. Several works [2,16,1,15] can be found in the literature to help a researcher assess the pros and cons of various techniques and take a decision accordingly. Although, a major challenge that remains relevant is how the scale of data impacts the clustering performance. Techniques like *hierarchical clustering* [5]

¹Corresponding Author: Ashutosh Karna, Knowledge Engineering and Machine Learning Group at Intelligent Data Science and Artificial Intelligence Research Centre, Universitat Politècnica de Catalunya, Catalonia, Spain; E-mail: ashutosh.karna@upc.edu

provide a tree-like graph, known as *dendrogram* that discloses the internal multivariate structure of the dataset and helps a researcher decide the appropriate number of partitions to make, but the quadratic time complexity of the algorithm prohibits its application on real-life datasets. Other techniques like *K-means* [7] do efficiently handle the large datasets but need a prior knowledge of K (number of clusters). Guha *et al.* proposed *CURE* (Clustering-Using-Representation) that scales hierarchical clustering to a large scale by incorporating a sampling strategy. In [13], Suman et al. proposed a new algorithm, *Bootstrap-CURE* that scales hierarchical clustering further by using several bootstrap samples in a *CURE* like strategy except the *shrinking* step. In this paper, the impact of *shrinking* and other related hyper-parameters are investigated on the overall clustering procedure in the context of the original dataset from 3D printing. The rest of the paper is structured as follows. Section 2 provides a summary of available research in this field, followed by a formal definition of the research problem in the section 3. Section 4 provides the proposed modification to original *Bootstrap-CURE* algorithm, followed by the methodology discussed in section 5. A summary of experiments conducted on a real-life dataset from 3D printing is discussed in section 6. The paper is finally concluded along with the discussion in section 7.

2. Literature Survey

As the size of the dataset increases, the increase in computational complexity makes it difficult to get clustering results in time, and thus traditional clustering methods are rendered impractical. Shirkhorshidi *et al.* [11] and Zerhari *et al.* [17] in their works, conducted a detailed investigation of clustering scenarios in context to big data and laid down the challenges in various methods. In both the works, the scale of data is rather suggested to be viewed in terms in *single vs multiple* machine-learning problem. For a single machine learning scenario, a small sample from the data is drawn from the larger dataset to use in clustering; while dimension-reduction techniques are used when the data is high-dimensional. Zhang *et al.* [18,19] introduced a novel clustering algorithm, called *BIRCH* (Balanced iterative reducing and clustering using hierarchies) to address the problem of processing large datasets with limited computing power. *BIRCH* introduces the concept of *Clustering-Factor* which is a triple (N, L_s, S_s) containing the number of items, linear sum, and squared sum of items in a subcluster respectively. *BIRCH* provides a two-step process, starting with one-pass scanning of the dataset and localized clusters are created. The localized clusters are expected to capture major patterns in the dataset and are further subject to another clustering. The method itself is based on top of the hierarchical clustering and scales it to big data.

McCallum *et al.* [8] proposed a canopy-based approach for clustering that is computationally cheap to estimate the distance matrix for large and high-dimensional datasets. The method efficiently divides the data into overlapping subsets, called *canopies* in the first stage, and then the distance measurements are computed in a common canopy.

CURE clustering [4] is another single machine learning technique that uses a set of representative points to estimate a cluster shape and shrink the representative points towards their respective cluster center.

In the earlier research, the authors [13,6] proposed a modification of *CURE* clustering using several bootstrap samples, however, the shrinking step from the original *CURE* algorithm is skipped.

Application of *shrinking* can be seen in several related works of clustering. In [14], Wang *et al.* proposed a new clustering algorithm based on local clustering that automates the discovery of clusters and right partitioning of data without any user input. In [3], Franti *et al.* proposed iterative-shrinking approach to clustering to obtain the suitable number of clusters. Shi *et al.* [10] uses shrinking as a data preprocessing step in high dimensional data clustering. In [9], Qian *et al.* proposed a modification of classical *CURE* clustering, called, *CURE-NS* which allows detecting non-spherical shaped clusters and ran experiments to compare the algorithm over the original *CURE* implementation.

3. Research Problem

Let us consider a multivariate numerical dataset, with the information about a set I of N , k -dimensional objects as i_1, i_2, \dots, i_N . The *Bootstrap-CURE* proposal [13] begins with drawing a small representative sample of ratio r from the original dataset, and divides it into S bootstrap samples of same size n_s without replacement. Each bootstrap sample is subject to hierarchical clustering individually and local clusters are obtained. A novel algorithm as described in [12] is used to deduce the number of clusters automatically. Further, let B_i represent i^{th} bootstrap sample and $c_{i,j}$ denote the local centroid of j^{th} cluster of i^{th} bootstrap sample. In the earlier research by the authors, the super-classification step involved applying a hierarchical clustering on all local centroids ($c_{i,j}, i \in 1, 2, \dots, S, j \in 1, 2, \dots, k_i$), where k_i is the number of clusters in i^{th} bootstrap sample and super centroids are obtained. In the second-pass of the algorithm, the unsampled points $N * (1 - r)$ are scanned and assigned to a super-cluster based on K-nearest-neighbor scheme.

The objective of this paper is to introduce the *shrinkage* step in the overall *Bootstrap-CURE* algorithm and assess its impact on the clustering results. In addition to *shrinkage*, following hyperparameters also play an important role in the quality of clustering and are evaluated as well.

1. **Sample Ratio (r):** Proportion of the original dataset drawn at random in step 1 of the Bootstrap CURE strategy.
2. **Number of Bootstrap samples (S):** Number of samples drawn without replacement from the initial sample obtained in step 1. Each bootstrap sample is of size $N * r / S$ and is individually clustered in step 2.
3. **Extreme or Reference points(q):** These refer to the points lying on the boundary of a cluster. They are identified as pairs of points with a bigger Euclidean distance between them, and capture the shape and extent of the cluster. The higher the value of q , better is the representation of the frontier of the cluster.
4. **Shrinkage (α):** It is a concept used in classical *CURE* definition. It implies moving extreme points towards the center of the cluster and using *shrunk* points as representatives of the cluster itself. Hence, the uncertainty associated with the frontier is reduced and robustness is gained. The higher the shrinkage rate, the closer are the reference points to the cluster centroid.

In this paper, the *Bootstrap-CURE* results with and without shrinking are compared.

4. Research Proposal

In the original *CURE* implementation, a set of extreme points in each cluster are selected as reference points which are then shrunk towards their respective cluster centers which help approximate the shape of each cluster in an unsupervised manner. In this research, the authors introduce and extend the concept of shrinking and the reference points to all the bootstrap samples.

The original *Bootstrap-CURE* algorithm is thus modified and summarised in the following steps:

1. **Initial Sampling phase:** Fix a sample ratio r and draw a random sample of size $N * r$ from the original dataset. Divide this random sample into S bootstrap samples, each of same size n_s , without replacement.
2. Subject each sample to hierarchical clustering (in this work with *Euclidean* distance and *Ward's* method), and obtain the S dendograms.
3. Use method proposed in [12] to determine the number of clusters automatically.
4. Compute the centroid of all clusters found in the previous step and build a final dataset with all local centroids, represented as $c_{i,l}, i \in 1, 2, \dots, S, l \in 1, 2, \dots, k_i$, where k_i is the number of clusters detected in i^{th} sample.
5. **Super-classification phase:** Apply a hierarchical clustering on the local centroids dataset and compute super-centroids of each super-class. Let each super-centroid be denoted by $C_g, g \in 1, 2, \dots, k_g$ where k_g is the number of clusters in super-clustering step. Each C_g represent a centroid of a set of local centroids ($c_{i,l}$) and let n_{c_g} be the size of g^{th} super-cluster.
6. Fix a percentage q and compute $n_{c_g} * q$ extreme points in each super-cluster. Let $x_{g,t}, g \in 1, 2, \dots, k_g, t \in 1, 2, \dots, n_{c_g} * q$ represent the t^{th} extreme point in g^{th} super-cluster.
7. **Shrinking phase:** Fix a shrinkage percentage, α and shrink each extreme point $x_{g,t}$ towards its super-centroid C_g . This is done by computing a synthetic point by shrinking the euclidean distance by p percent between the extreme point and the super-centroid.
8. Trace all points from the sampled data belonging to each local centroid and in turn, to each super-centroid.
9. **Allocation phase:** For all $N * (1 - r)$ points which are not part of the original sample, assign the super-cluster based on the nearest shrunk extreme point.

In order to evaluate the impact of various hyperparameters (as mentioned in section 3), the original dataset is subject to *Bootstrap-CURE* with and without *shrinkage* respectively and the class assignments of each point in both the cases is tracked and assessed how the two approaches differ in the clustering results. The coincidences are computed for each cluster individually and the average coincidence rate (*ACR*) is obtained.

An illustrative example is discussed in section 6 using a real-life dataset from 3D printers.

5. Research Methodology

In this research, the authors are introducing the *Shrinkage* parameter during the super-clustering step of *Bootstrap-CURE* clustering. The objective of the study is to see how

various hyperparameters impact the clustering assignment in *Bootstrap-CURE* process. Hence, for each experiment, a pre-decided combination of hyperparameters is used and the data is subject to *Bootstrap-CURE* clustering with both *with* and *without* shrinkage, and the class assignment is studied.

Following the terminology defined in section 3, the general *Bootstrap-CURE* clustering can be formulated as the following:

$$\mathcal{B}_\alpha = \phi(r, S, \alpha, q); r \in [0, 1], S \geq 2, 0 \leq \alpha < 1, 0 < q \leq 1 \quad (1)$$

where α denotes the shrinkage.

The Eq 1 reduces to *Bootstrap-CURE clustering without shrinkage* when $\alpha = 0$, and let this be denoted by \mathcal{B}_0 . Let k_α and k_0 denote the number of clusters obtained by using \mathcal{B}_α and \mathcal{B}_0 clustering algorithms respectively. Further, let A_k denote the coincidence for k^{th} cluster respectively and \bar{A} denote the average coincidence rate (ACR) for overall data.

6. Application

The dataset used in the earlier research [13] has been continued for the current experiments. The data represent a collection of time-series-based sensor data from eight anonymous 3D printers with over 300 printing jobs. All experiments have been conducted on a GPU-enabled, four-core processor windows computer with 32 GB memory and *Python 3.6* has been used throughout for data analysis. The final working data after preprocessing contains 46821 records for 41 features. A list of features can be seen in the previous paper [13].

6.1. *Bootstrap-CURE Clustering*

In the earlier research [13], the authors proposed and conducted *Bootstrap-CURE* clustering on several samples of varying sizes drawn from the original dataset and subject them to hierarchical, *CURE* and *Bootstrap-CURE* clustering. While the number of clusters discovered remains the same (four clusters), the *Bootstrap-CURE* evidently showed a rapid decrease in the computation time as the dataset size increased. A summary of experimental results can be seen in [13].

6.2. *Bootstrap-CURE with Shrinkage*

In this research, the authors have modified the original *Bootstrap-CURE* algorithm with a new step, called *shrinking* added just before the final cluster-assignment stage. The cluster assignment before and after shrinkage is then tracked. Table 1 shows a schema of contingency matrix to compare the clustering assignments by the original (*without shrinkage*) and modified (*with shrinkage*) *Bootstrap-CURE* clustering algorithms, with O_{ij} representing the number of items classified into i^{th} cluster of original and j^{th} cluster of the modified *Bootstrap-CURE* algorithm. Further, coincidence rate for each individual cluster class ($k=1,2,\dots,k_0$) is computed.

		Bootstrap-CURE with shrinkage			
		1	2	...	k_α
Bootstrap-CURE without shrinkage	1	O_{11}	O_{12}	...	O_{1k_α}
	2	O_{21}	O_{22}	...	O_{2k_α}

	k_0	O_{k_01}	O_{k_02}	...	$O_{k_0k_\alpha}$

Table 1. Contingency matrix layout for *Bootstrap-CURE*

6.3. Impact of hyperparameters on *Bootstrap-CURE*

As discussed in section 5, the *Bootstrap-CURE* algorithm depends on 4 hyperparameters. For this research, the following range of values for these hyper-parameters have been tested.

- Initial sample ratio (r): $r \in [0.10, 0.20, 0.50, 0.75]$
- Number of bootstrap samples (S): $S \in [5, 10, 15, 20]$
- Percentage of extreme/reference points (q): $q \in [0.02, 0.05, 0.10, 0.20]$
- Percentage of shrinkage (α): $\alpha \in [0.05, 0.10, 0.15, 0.20]$

A total of 256 experiments were conducted for every possible combination of the hyperparameters defined above. In the following, the main effect of each hyperparameter on the ACR is summarized.

6.3.1. Impact of Initial Sample Ratio (r)

The size of the initial sample drawn from the original data directly impacts the clustering quality. Fig 1 shows the impact of other hyperparameters on ACR conditioned to sample ratio. It is easily evident that a higher sample ratio ($r \geq 0.5$) yields higher ACR for all hyper-parameters. Also, given a certain sample ratio, the ACR seems to be constant per shrinkage level, and increases with increase in percentage of extreme points. However, the trend is not clear regarding the number of bootstrap samples.

6.3.2. Impact of Number of Bootstrap samples (S)

In theory, the increase in the number of bootstrap samples improves the computational time of clustering as it allows parallel execution. However, as a hyperparameter, the number of bootstrap samples does not increase the ACR by itself. Conditioning to a given number of bootstrap samples, higher sample ratio has clear impact on better ACR but does not show any impact on other hyperparameters like shrinkage level or percent of extreme points as shown in Fig 2.

6.3.3. Impact of Shrinking Percentage (α)

Four levels of shrinkage have been considered for the experiments. In line with what was observed in previous graphs, the shrinkage level is not determining an average improvement on ACR. When conditioning to a given shrinkage level, one can see that the ACR increases with the sample ratio, and a very weak improvement is observed with the increase in percentage of extreme points. The figure has similar pattern as fig 2

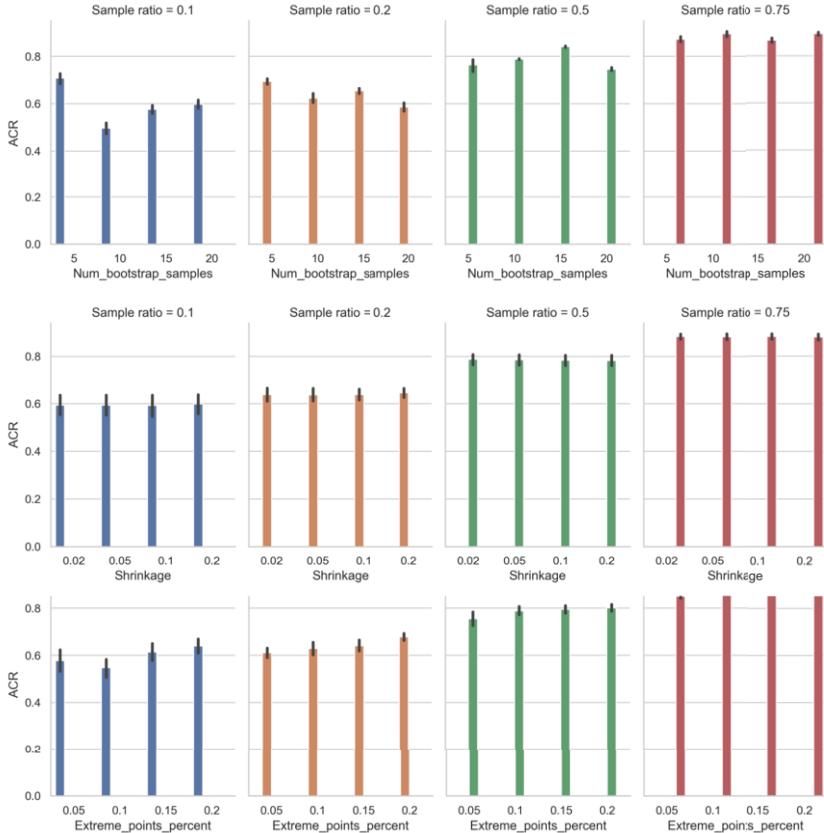


Figure 1. Impact of sample ratio

6.3.4. Impact of Percentage of extreme points (q)

The reference points are used during the super classification step of the *Bootstrap-CURE*. Fig 1 shows that the percent of reference points do not affect the average ACR by itself. Conditioning to a given percent of extreme points, the ACR increases with the sample ratio. The figure has similar pattern as 2.

7. Discussion and Conclusion

In this paper the original *Bootstrap-CURE* algorithm is modified by introducing shrinking after the super-classification step and consequently, the clustering assignments are evaluated both with and without shrinkage. The research is an attempt to discover the intrinsic nature of hyper-parameters and their interactions in a *Bootstrap-CURE* clustering procedure.

In general, the increase in *initial-sample-ratio* always increases the quality of the clustering. However, a higher sample ratio also implies a higher computational cost as the size of the data to be clustered increases. A lower sample ratio reduces the computation cost but the representation of the data adversely affects the clustering quality. A bal-

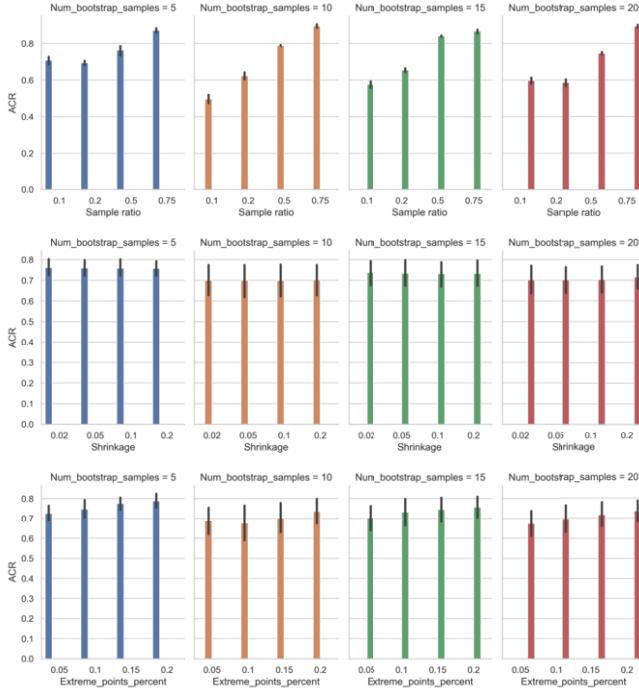


Figure 2. Impact of number of bootstrap samples

ance between sample ratio and quality (ACR) should be sought by the researchers. In the earlier research work, the authors used $r = 0.5$ as the initial sample ratio and this seems reasonable to the current experimental results. The number of bootstrap samples is another important factor that impacts the computational cost in *Bootstrap-CURE*. The experiments show when the number of bootstrap samples is fewer, the sample ratio can be increased to achieve better clustering quality. The application of shrinkage, however, is quite agnostic to the quality of clusters, and it seems that this is due to the topology of the analyzed data. This dataset has a sufficiently strong structure with fewer observations in the areas of inter-clusters frontiers that might oscillate between one cluster to another depending on the shrinkage level. When the percentage of reference points increases, we get a better representation of the cluster shape and ACR shows improvement.

The experimental results indicate that the introduction of shrinkage does not impact the clustering quality for the tested domain. However, the number of bootstrap samples and the initial sample ratio play a rather more important role in deciding the overall quality, as well as the number of extreme points.

In the future lines, further experimentation is planned with a synthetic dataset with different topologies, so as to verify how the shrinkage percentage impacts ACR in different kinds of data structures. In theory, shrinkage improves the robustness of clusters as it manages the outliers better. Further analysis is being conducted to check the outliers' structure of the target dataset. From the computational point of view, introducing shrinkage does not show an observable impact as it can be implemented like a scaling operation on observation vectors with a fraction of milliseconds in computational costs.

Mathematical optimization of the hyperparameters has not been part of this research but would be very useful in the upcoming research in order to finetune the clustering results.

This research directly fits into a bigger project that aims at developing an intelligent decision support system for enterprise-scale 3D printers and the clustering procedure is aimed at the automatic discovery of patterns without human intervention. In the future line of research, the authors also aim to build the layout of decision support system and investigate how the clustering scheme fits into the overall strategy of detecting different operating modes in the machine during the runtime.

References

- [1] Sabhia Firdaus and Md Uddin. A survey on clustering algorithms and complexity analysis. *International Journal of Computer Science Issues (IJCSI)*, 12(2):62, 2015.
- [2] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.
- [3] Pasi Fräntti and Olli Virmajoki. Iterative shrinking method for clustering problems. *Pattern Recognition*, 39(5):761–775, 2006.
- [4] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. *Information Systems*, 26(1):35–58, 3 2001.
- [5] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
- [6] Ashutosh Karna and Karina Gibert. Automatic identification of the number of clusters in hierarchical clustering. *Neural Computing and Applications*, pages 1–16, 2021.
- [7] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.
- [8] Andrew McCallum, Kamal Nigam, and Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, 2000.
- [9] Yun-Tao Qian, Qing-Song Shi, and Qi Wang. Cure-ns: A hierarchical clustering algorithm with new shrinking scheme. In *Proceedings. International Conference on Machine Learning and Cybernetics*, volume 2, pages 895–899. IEEE, 2002.
- [10] Yong Shi, Yuqing Song, and Aidong Zhang. A shrinking-based clustering approach for multidimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 17(10):1389–1403, 2005.
- [11] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, and Tutut Herawan. Big data clustering: a review. In *International conference on computational science and its applications*, pages 707–720. Springer, 2014.
- [12] Shikha SUMAN, Ashutosh KARNA, and Karina GIBERT. Towards expert-inspired automatic criterion to cut a dendrogram for real-industrial applications. *ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT*, page 235, 2021.
- [13] Shikha Suman, Ashutosh Karna, and Karina Gibert. Bootstrap-cure: A novel clustering approach for sensor data—an application to 3d printing industry. *Applied Sciences*, 12(4):2191, 2022.
- [14] Xiaogang Wang, Weiliang Qiu, and Ruben H Zamar. Clues: A non-parametric clustering method based on local shrinking. *Computational Statistics & Data Analysis*, 52(1):286–298, 2007.
- [15] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.
- [16] Mohamed Zait and Hammou Messatfa. A comparative study of clustering methods. *Future Generation Computer Systems*, 13(2-3):149–159, 1997.
- [17] Btissam Zerhari, Ayoub Ait Lahcen, and Salma Mouline. Big data clustering: Algorithms and challenges. In *Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA’15)*, 2015.
- [18] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.
- [19] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1(2):141–182, 1997.

Adapting a Fuzzy Random Forest for Ordinal Multi-Class Classification

Jordi PASCUAL-FONTANILLES^{a,1}, Lenka LHOTSKA^b Antonio MORENO^a, and Aida VALLS^a

^aITAKA, Dept. Enginyeria Informàtica i Matemàtiques

Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

^bCzech Institute of Informatics, Robotics and Cybernetics

Czech Technical University in Prague, Czechia

Abstract. Fuzzy Random Forests are well-known Machine Learning ensemble methods. They combine the outputs of multiple Fuzzy Decision Trees to improve the classification performance. Moreover, they can deal with data uncertainty and imprecision thanks to the use of fuzzy logic. Although many classification tasks are binary, in some situations we face the problem of classifying data into a set of ordered categories. This is a particular case of multi-class classification where the order between the classes is relevant, for example in medical diagnosis to detect the severity of a disease. In this paper, we explain how a binary Fuzzy Random Forest may be adapted to deal with ordinal classification. The work is focused on the prediction stage, not on the construction of the fuzzy trees. When a new instance arrives, the rules activation is done with the usual fuzzy operators, but the aggregation of the outputs given by the different rules and trees has been redefined. In particular, we present a procedure for managing the conflicting cases where different classes are predicted with similar support. The support of the classes is calculated using the OWA operator that permits to model the concept of majority agreement.

Keywords. Fuzzy Random Forest, Multi-class ordinal classification, Ensemble classifiers, OWA operator

1. Introduction

A Fuzzy Random Forest (FRF) is an extension of Random Forests which makes use of fuzzy logic. This addition allows them to manage uncertainty and imprecision of the data. It is composed by a set of Fuzzy Decision Trees (FDT), which can be constructed using several algorithms. This paper continues our previous work on the construction and use of FRF for binary classification in health care. Our construction method is based on Yuan and Shaw's induction algorithm [1] with some extensions presented in [2]. The algorithm has two parameters: α is the threshold indicating the minimum membership degree considered during inference, and β is the minimum truth level required to generate a new rule. In the FRF model presented in [2], the classification has 2 steps. The first step is at the FDT level, where the predictions of the rules are aggregated to decide the output class given by the tree. The second step consists on aggregating the outputs of

¹Corresponding Author: Jordi Pascual-Fontanilles. E-mail: jordi.pascual@urv.cat

all the FDTs to make the final binary class assignment. Two parameters were introduced in these steps (δ_1 and δ_2) to allow the assignment of the *Unknown* category when the system is not sure about which of the two classes is the winner.

In ordinal multi-class decision problems we must assign a class to an instance from a set of k ordered possibilities $C = \{Class_0, Class_1, \dots, Class_{k-1}\}$, where $k > 2$. Depending on the problem, the order can be of increasing or decreasing preference, also called Gain or Cost. For example, in medical diagnosis, the usual order goes from the best to the worst medical conditions, so that $Class_0$ has the healthy people and the greater the class index, the worse is the disease level.

In this paper, we adapt the 2-step classification process of FRF for the case of ordinal multi-class decision problems. In Section 2 the first classification step is adapted. Section 3 explains the modifications on the second classification step. Section 4 shows experimental results. Finally, Section 5 gives the conclusions and future work.

2. Fusion in a Fuzzy Decision Tree

In a FDT we have a hierarchical structure with r branches from the root node to the leaves. Each branch corresponds to a different fuzzy rule with one or more premises consisting of linguistic variables defined on fuzzy sets. When rules are learned automatically from examples, in the leave nodes we can store a rule support value for each possible class.

When a new instance is classified, each rule provides a decision support value for each of the available classes, obtained from the product of the rule premises activation and the rule support for each class. Therefore, for each class $Class_i$ we obtain a tuple with r decision support values, one for each rule: $D_{i,1}, D_{i,2}, \dots, D_{i,r}$.

To decide which is the final class assigned to the example, we must take into account the overall support received by each class. To merge the values provided by all rules, in [3] we analysed several aggregation operators, and we proposed the use of the Choquet fuzzy integral, with a fuzzy measure based on the distorted probability. Before continuing, the support value is normalized using the truth level threshold β used for constructing the rules. The maximum support allowed is 1. So, for the i -th class, we have the support calculated using Eq. (1).

$$D_i^N = \min \left(1, \frac{\text{ChoquetIntegral}(D_{i,1}, D_{i,2}, \dots, D_{i,r})}{\beta} \right) \quad (1)$$

In our previous work [2], a threshold value δ_1 was introduced for binary classification, to determine if the FDT had a clear consensus on determining the winner class. To avoid mistakes, the label *Unknown* was introduced. When the difference between the two decision support values is lower than δ_1 , we assume that the FDT is not sure and, hence, the label *Unknown* is assigned. The use of δ_1 is maintained in this proposal for multi-class classification, but, because of the multiple classes, the method has been adapted. In binary classification, δ_1 was just compared with the difference of the support to the two classes. To use this threshold to the case of multiple classes, we propose the following strategy.

We order decreasingly the set $C^* = \{C \cup Unknown\}$, according to the normalized decision support values D_i^N . Let us consider that C_a is the category with the highest decision support and C_b is the second most-supported category. The result of the analysis of the j -th FDT is a tuple with the predicted class and its support (P_j, S_j) . The final prediction associated with the FDT is chosen between these two categories, as described by Eq. (2). One of the strengths of an ensemble is the diversity of models composing it. An unknown prediction is preferred when the model has not a unique preferred class, which is better than making an incorrect prediction. The next section explains how the ensemble aggregates the predictions of the different trees to get the final decision.

$$(P_j, S_j) = \begin{cases} (Unknown, 0), & \text{if } D_a^N - D_b^N < \delta_1 \\ (C_a, D_a^N), & \text{otherwise} \end{cases} \quad (2)$$

3. Fusion in a Fuzzy Random Forest

Once all the FDTs have made a prediction about the output class, all the predictions on the ensemble are aggregated to decide the final class assignment and its support. An ensemble formed by n FDTs has a set of n predicted classes, each with a support value: $(P_1, S_1), (P_2, S_2), \dots, (P_n, S_n)$

In the following subsections, we explain the proposal to aggregate all the predictions of the ensemble on the multi-class case. Its main elements are a weighted voting, some heuristics for the final class assignment and an OWA-based decision support score.

3.1. Weighted Voting

A voting process is used to find the consensus class from the ensemble of different FDTs. Each FDT has a weight assigned to it, which represents its prediction quality. It is computed using the out-of-bag examples on the training phase. A quality metric has to be properly chosen to represent the overall quality of each FDT.

To aggregate all the predictions of a class through weighted voting, the weights of the trees that predicted it are summed, Eq. (3). As a result, each of the classes obtains a voting value v_i , which is used to decide the final prediction of the ensemble, as explained in the next subsection. Thus,

$$v_i = \sum_{j \in I} w_j \quad (3)$$

, where w_j is the weight assigned to the j -th tree of the set $I = \{t \mid P_t = Class_i\}$.

In the previous binary approach, we tested several metrics for weighting the trees. An average accuracy balancing sensitivity (2/3) and specificity (1/3) was used [4]. This balanced accuracy is specially useful in domains such as the medical one, in which a good sensitivity is a priority in order to avoid false negatives.

For the case of multi-class problems, the most appropriate and usual quality measures are $F1$ (balancing precision and recall) and the Weighted Cohen's Kappa κ . If we take into account the order between the classes, then κ is the best performance index,

because it allows to define different penalization for mistakes between classes depending on the distance between them [5]. For this reason, we propose to use κ in the weighting process of FRF on ordinal multi-class classification.

3.2. Final Class Assignment

In binary classification, the final class assignment is made similarly to the selection of the class in a FDT, with the comparison of the support obtained by the two classes, in this case, the votes. In [2] the constant parameter δ_2 was introduced to detect the cases where the difference in votes between the two classes is not significant. In that case, when the difference in votes is lower than δ_2 , the *Unknown* category is returned by the classification model to avoid mistakes. So, the final class A was obtained as follows:

$$A = \begin{cases} \text{Unknown}, & \text{if } v_0 - v_1 < \delta_2 \\ C_a, & \text{otherwise (Class}_{0/1}\text{ with higher support)} \end{cases} \quad (4)$$

With multiple classes, we will take C_a and C_b again as the first and second most voted classes respectively. In this paper, for the multi-class proposal, δ_2 is preserved, but it is defined as a function depending also on the difference between the two most voted classes, according to their position in the ordered set of possible categories C . Let us define Δv_{ab} as the normalized difference of votes between C_a and C_b , Eq. (5).

$$\Delta v_{ab} = \frac{v_a - v_b}{\sum_{i \in C^*} v_i} \quad (5)$$

With this normalization, the δ_2 threshold is now defined in two parts, Eq. (6). A first constant part $d \in [0, 0.5]$, which is the minimum difference in votes that permits to distinguish the support of the classes and make a class assignment. In addition, the separation between the classes in the ordered scale C is also relevant to define when a difference in votes is important or not. It is not the same choosing between consecutive classes than between extreme classes in C . For that reason, the second part of δ_2 is given by the square of the difference between the positions of the classes. The value is limited to 0.5, for all the cases where the most voted class has more than half of the total votes.

Formally, let us define $\text{index} : \text{class} \rightarrow [0, k - 1]$ as the function that returns the position of a given class in the ordered set C , and the distance between classes as $\text{dist}(C_a, C_b) = |\text{index}(C_a) - \text{index}(C_b)|$. Then, the definition of δ_2 is the following:

$$\delta_2 = \begin{cases} d, & \text{if } C_a = \text{Unk} \text{ or } C_b = \text{Unk} \\ \min(0.5, d + \frac{\text{dist}(C_a, C_b)^2}{100}), & \text{otherwise} \end{cases} \quad (6)$$

Using this new δ_2 definition in Eq. (4) is a quite conservative approach that generates many assignments to the *Unknown* category. To avoid that the classifier does not provide an answer in too many cases, the following heuristics for assignments are proposed:

- H1: When one of the two most voted classes is the *Unknown* category and the difference in votes is small, then the model returns the class $C_n \neq Unknown$. However, if the difference is large enough, then the model returns the most voted class. It may be *Unknown* or the other one.
- H2: If the two classes are not unknown and the difference in votes is large enough, the most voted class must be the output of the classification model.
- H3: In the cases where the two classes are not unknown and the number of votes is similar, $\Delta v_{ab} < \delta_2$, two options are considered, depending on the distance of the classes in the ordered set C . If there is a big distance between positions of the classes in the ordered set C , the ensemble is considered not being certain about the prediction, and the label *Unknown* is assigned. In the case of close classes in C , the selected class is the one with a higher index in this ordered set. The distance threshold is based on the number of classes k .

These heuristics are formalized in the following equation:

$$A = \begin{cases} C_n, & \text{if } (C_a = Unk \text{ or } C_b = Unk) \text{ and } \Delta v_{ab} < \delta_2 \\ C_a, & \text{if } (C_a = Unk \text{ or } C_b = Unk) \text{ and } \Delta v_{ab} \geq \delta_2 \\ Unk, & \text{if } C_a \neq Unk \text{ and } C_b \neq Unk \text{ and } \Delta v_{ab} < \delta_2 \text{ and } dist(C_a, C_b) \geq \lfloor \frac{k}{2} \rfloor \\ C_m, & \text{if } C_a \neq Unk \text{ and } C_b \neq Unk \text{ and } \Delta v_{ab} < \delta_2 \text{ and } dist(C_a, C_b) < \lfloor \frac{k}{2} \rfloor \\ C_a, & \text{if } C_a \neq Unk \text{ and } C_b \neq Unk \text{ and } \Delta v_{ab} \geq \delta_2 \end{cases} \quad (7)$$

, where $C_n \neq Unknown, n \in \{a, b\}$, and $m = \max(index(C_a), index(C_b))$.

Notice that we assumed a minimization goal, where wrong classification to less severe classes is not desired. If the goal is maximization, then m should be the minimum.

3.3. Final Decision Support

Together with the predicted class, A , the FRF calculates a decision support value of the prediction. This support is obtained from the corresponding support values given by each decision tree. An arithmetic average is usually used as aggregation operation. In this work, we propose the Ordered Weighted Average (OWA) to perform the aggregation [6]. OWA is a parameterized operator that permits to make a conjunctive or disjunctive aggregation. The polarity of the operation is defined with a set of weights assigned to the input values according to their position after their reordering. Having a set of support values S_j obtained with Eq. (2) for each tree, and having a weight for each position $w_i, i = 1..n$. The result F is obtained with Eq. (8), where $S_{\sigma(i)} < S_{\sigma(i+1)}$.

In a FRF the number of trees, n , is usually large (i.e. hundreds), but only a subset of the trees corresponds to the final class A . Given the randomness in the selection of attributes, some of these trees may produce low support values. However, if a sufficient number of trees, $m << n$, is highly supporting the selected class, the confidence about this class should be high (disjunctive policy). The weighting vector, where $\sum_{i=1}^n w_i = 1$, has been defined with weights that decrease, Eq. (8).

$$F = \sum_{i=1}^n w_i S_{\sigma(i)}, \text{ where } w_i = \frac{i}{\sum_{j=n-m}^n j}, \text{ for } i \in [n-m, n], \text{ and } w_i = 0 \text{ otherwise} \quad (8)$$

4. Experiments

4.1. Dataset

The experiments will be done with the diabetic retinopathy (DR) risk detection problem. In the last years, we developed the RETIPROGRAM system [7]. It is based on a binary FRF classifier, which proved to give the best results for this problem [2]. The model considers 9 different attributes (6 numerical and 3 categorical) to distinguish between the positive and the negative class. The numerical attributes were fuzzified with the ophthalmologists expertise, defining appropriate linguistic labels [7].

To test the ordinal multi-class proposal, we used a dataset from a private regional hospital. The dataset includes real data from 2084 diabetic patients. The ETDRS standard classification is considered for the target DR attribute [8]. They are ordered from lowest to highest degree of DR, $C = \{NoDR, Mild, Moderate, Severe\}$. The data has been split in two different datasets, training (80%) and testing (20%). Table 1 shows the distribution of the data among the target attribute classes, which has a large imbalance towards the first class, *NoDR*.

Table 1. Diabetic retinopathy data distribution

	Training	Testing	Total
NoDR	1394 (83.6%)	349 (83.7%)	1743
Mild	191 (11.5%)	48 (11.5%)	239
Moderate	58 (3.5%)	14 (3.4%)	72
Severe	24 (1.4%)	6 (1.4%)	30
Total	1667	417	2084

4.2. Study of the Weights of FDTs in the Voting Stage

From the different contributions presented in this paper for the case of ordinal multi-class assignments with FRF, we start by testing the effect of using the κ index instead of Accuracy to give a weight to each of the trees in Eq. (2). We compare 3 versions of the FRF classification algorithm:

1. Base algorithm: it does not consider the category *Unknown*, so that we always classify an instance to one of the output classes. Hence, $\delta_1 = 0$ and $\delta_2 = 0$.
2. Base- δ algorithm: it takes into account situations where two classes have similar conditions and then the answer is unknown, to try to avoid mistakes.
3. New- δ algorithm: it corresponds to the new procedure explained in this paper.

We will denote as FN (False Negatives) to the examples where the model predicts as a class lower than the real (i.e. underestimation or type-II error). Similarly, we call FP (False Positives) when the predicted class is higher than the real one (i.e. overestimation or type-I error). FNs are a kind of error non desirable in medical diagnosis, because the system does not detect the real risk for the health of the person.

We have defined the Base version as the model to improve, as it makes too many mistakes. The confusion matrix in Table 2 shows the results of the Base version. For example, in *Mild*, from the total of 48 patients, we have 15 classified to *NoDR* (FN=31%). Similarly, there is a 28% of FN in *Moderate* and 33% in *Severe*.

Table 2. Base method confusion matrix

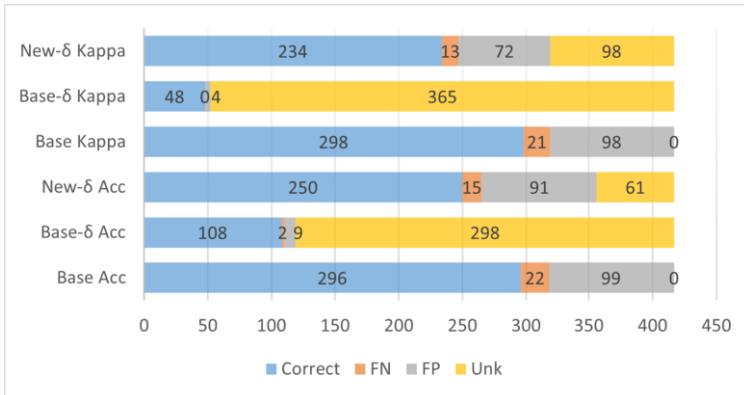
Real/Predicted	NoDR	Mild	Moderate	Severe
NoDR	278	30	21	20
Mild	15	13	8	12
Moderate	2	2	3	7
Severe	2	0	0	4

Table 3 compares the results of using Accuracy or κ as the quality metric in the weighted voting. The 3 versions of the algorithm are compared on the Accuracy (Acc), Accuracy including unknowns as errors (Acc Unk), and Kappa index. Thresholds used are $\delta_1 = 0.1$ and $\delta_2 = 0.25$, which have proven through empirical testing to be the best values. An in depth analysis of δ_2 is shown in subsection 4.3.

Table 3. Comparison of two weighted voting quality metrics

Method/Weight	Accuracy Weight			κ Weight		
	Acc	Acc Unk	Kappa	Acc	Acc Unk	Kappa
Base	0.71	0.71	0.34	0.715	0.715	0.345
Base-δ	0.908	0.259	0.529	0.923	0.115	0.509
New-δ	0.702	0.6	0.312	0.734	0.561	0.318

Better quality values are obtained using κ as weights of the trees. The accuracy index increases in the 3 versions. Weighted Kappa index is maintained to a similar level. Acc Unk metric decreases a bit, meaning the κ Weight produces more unknown predictions than Accuracy Weight. To further analyse the weight selection in the voting stage, Figure 1 shows the distribution of correct, incorrect and unknown predictions.

**Figure 1.** Distribution of correct, incorrect and unknown class assignments for different voting weights

The Base- δ algorithm does not perform appropriately. It has very few errors, but there are too many unknown predictions. In contrast, the New- δ algorithm is able to reduce the incorrect predictions by introducing a moderate amount of unknowns. Comparing Accuracy and κ in the New- δ algorithm, even though κ has less correct predictions, the amount of incorrect predictions is also smaller. We consider κ to have better results because of its more conservative results on unclear cases. Moreover, it obtains a better global accuracy.

4.3. Study of δ_2 for Class Assignment in Ordinal FRF

We studied the effect of using different d values to compute δ_2 for the final class assignment. Figure 2 shows the effect of d on the distribution of predictions among correct, incorrect and unknown. As expected, the higher d , the lower the number of unknown predictions. This is due mainly to decreasing the number of cases that enter to the second condition in Eq. (7). Accordingly, correct and incorrect assignments increase as the amount of unknowns decreases. We can see that the d parameter allows modelling the trade-off between correct, incorrect and unknown predictions. For the DR risk assessment problem, $d = 0.25$ was chosen for its balance of reducing incorrect predictions while not increasing unknowns and reducing correct predictions in excess. In other domains, δ_2 can be adapted according to the problem being solved, and the implications of miss-classifications.

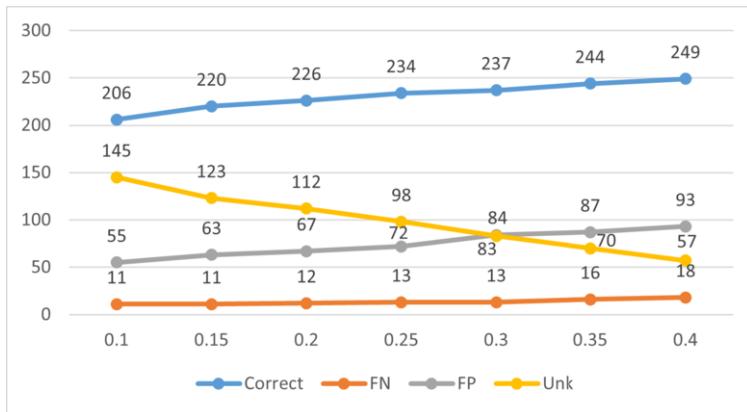


Figure 2. Distribution of correct, incorrect and unknown class assignments for different d values

4.4. Study of the Heuristics for Class Assignment in Ordinal FRF

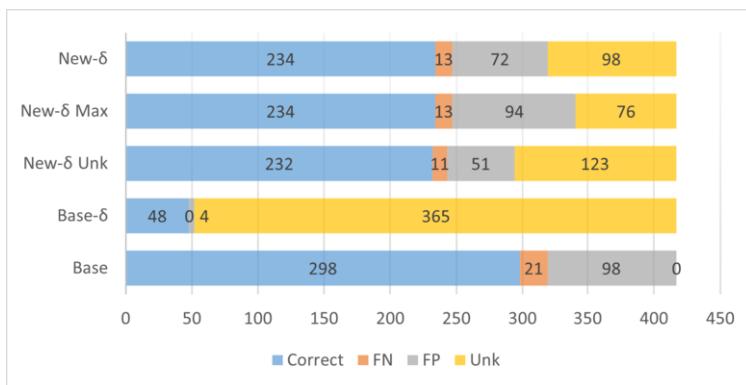
To study the effects of the proposed heuristics, the algorithm versions explained in 4.2 have been tested with two additional versions, Table 4. The additional versions differ in heuristic H3, which considers cases with the two most voted classes not being *Unknown* and a similar number of votes, $\Delta v_{ab} < \delta_2$. We eliminate the condition about the distance of the classes in the ordered set C , instead, a predetermined label is assigned. New- δ -Unk assigns *Unknown*, whereas New- δ -Max assigns the class with the higher index in C .

The performance improvement can be more clearly seen in the distribution of correct, incorrect and unknown assignments in Figure 3. Comparing New- δ -Unk and New- δ -Max with New- δ , we can conclude that by taking into account the distance between the majority classes, we can balance the number of errors and unknown assignments. Even though New- δ -Unk is the version with fewer errors, it is not the preferred version, as it could lead to having too many predictions assigned to *Unknown*. In the case of New- δ -Max, we can see on the FP the effect of classifying to the class with the higher index when the FRF does not have a clear consensus towards one class. By merging

Table 4. Comparison of different versions of the method

	Accuracy	Acc Unk	Kappa
Base	0.715	0.715	0.345
Base-δ	0.923	0.115	0.509
New-δ-Unk	0.789	0.556	0.38
New-δ-Max	0.686	0.561	0.227
New-δ	0.734	0.561	0.318

both versions depending on the distance between classes, the amount of unknowns can be balanced while prioritising the classes with higher indexes. This behaviour is desired in ordinal cases such as the DR risk assessment, where a FN would have much worse consequences than a FP.

**Figure 3.** Distribution of correct, incorrect and unknown assignments in different versions of the FRF

4.5. Study of OWA for Final Decision Support Averaging

To study the effect of a disjunctive OWA in the aggregation of the decision support, it has been compared to an arithmetic average aggregation (AA), Table 5. Experiments have been performed with $n = 100$ number of trees and $m = \frac{n}{3}$ as the minimum number of trees supporting the selected class. The decision support values obtained from the test dataset, which can range in $[0, 1]$, have been split in three intervals, to indicate three levels of confidence on the answer given to the user. For each of them, the number of correct predictions is counted. We consider we should not have a low decision support in cases where a sufficient number of trees is sure about the prediction. This is the result achieved by OWA. The number of predictions in the higher intervals is greater than using an arithmetic average. As a consequence, the percentage of correct predictions in the higher interval is also greater. With AA the user has more uncertain answers, which in medicine are cases that require additional attention by the doctors, spending time and resources. So, OWA operator is recommended.

Table 5. Decision support values with AA and disjunctive OWA

	Average			OWA		
	[0, 0.5]	(0.5, 0.75]	(0.75, 1]	[0, 0.5]	(0.5, 0.75]	(0.75, 1]
Total	44	205	70	5	113	201
# correct	37	149	48	5	80	149
% correct	84 %	73 %	69 %	100 %	71 %	74 %

5. Conclusions and future work

In this paper, we presented an adaptation of a binary FRF model for ordered multi-class classification. We have focused on the 2 steps of the prediction stage, and we have re-defined the procedure to manage conflicting cases. The different contributions presented have been studied on a DR dataset. From the results we conclude that: κ index works better than accuracy for weighting the trees; we can model the trade-off between predictions and unknowns using δ_2 ; the proposed heuristics balance the number of unknowns while prioritising classes with higher indexes, which is desired in medical applications. Finally, OWA gives an appropriate confidence value on the class assigned by the FRF.

As future work, we should test the proposed method with other datasets to confirm the observations. Then, we plan to test the method with other aggregation operators, as well as to study how it could make use of the dynamic updating method proposed in [4].

Acknowledgements

This work is funded by projects PI21/00064 and PI18/00169 (Instituto de Salud Carlos III & FEDER funds), and 2020PFR-B2-61 (URV). The first author has a pre-doctoral FI grant (2021 FI_B 00139) from Generalitat de Catalunya and Fons Social Europeu. Prof. Lhotska is supported by the resources of the Czech Technical University in Prague.

References

- [1] Yuan Y, Shaw, MJ. Induction of fuzzy decision trees. *Fuzzy Sets and systems*. 1995;69(2):125-139.
- [2] Saleh E, Błaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernández S, et al. Learning ensemble classifiers for diabetic retinopathy assessment. *Artif Intell Med*. 2018;85:50–63.
- [3] Saleh E, Valls A, Moreno A, Romero-Aroca P, Torra V, Bustince H. Learning Fuzzy Measures for Aggregation in Fuzzy Rule-Based Models. In: *Lecture Notes in Computer Science*. Springer Verlag; 2018; p. 114–27.
- [4] Pascual-Fontanilles J, Valls A, Moreno A, Romero-Aroca P. Iterative Update of a Random Forest Classifier for Diabetic Retinopathy. *Frontiers in Artificial Intelligence and Applications*. 2021 Oct 14;339:207–16.
- [5] De La Torre J, Puig D, Valls A. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*. 2017;105:144–54.
- [6] Yager RR. On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decisionmaking. *IEEE Trans Syst Man Cybern*. 1988;18(1):183–90.
- [7] Romero-Aroca P, Valls A, Moreno A, Sagarraga-Alamo R, Basora-Gallisa J, Saleh E, et al. A Clinical Decision Support System for Diabetic Retinopathy Screening: Creating a Clinical Support Application. *Telemecodine and e-Health*. 2019 Jan 1;25(1):31–40.
- [8] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110(9):1677–82.

On Flows of Neural Ordinary Differential Equations That Are Solutions of Lotka-Volterra Dynamical Systems

Argimiro ARRATIA^{a,1}, Carlos ORTIZ^b and Marcel ROMANÍ^a

^aSoft Computing Research Group (SOCO)

at Intelligent Data Science and Artificial Intelligence Research Center
Department of Computer Sciences,

Polytechnical University of Catalonia, Barcelona, Spain.

argimiro@cs.upc.edu, marcel.romani@estudiantat.upc.edu

^bDept. of Computer Science and Mathematics,

Arcadia University, Glenside, PA, U.S.A.

ortiz@arcadia.edu

Abstract. Neural Ordinary Differential Equations (NODE) have emerged as a novel approach to deep learning, where instead of specifying a discrete sequence of hidden layers, it parameterizes the derivative of the hidden state using a neural network [1]. The solution to the underlying dynamical system is a flow, and various works have explored the universality of flows, in the sense of being able to approximate any analytical function. In this paper we present preliminary work aimed at identifying families of systems of ordinary differential equations (SODE) that are universal, in the sense that they encompass most of the systems of differential equations that appear in practice. Once one of these (candidate) universal SODEs is found, we define a process that generates a family of NODEs whose flows are precisely the solutions of the universal SODEs found above. The candidate universal SODE family that we present here is the generalized Lotka-Volterra (LV) families of differential equations. We present the NODE models built upon this LV systems and a description of their appropriate flows and some preliminary implementations of this process.

Keywords. Neural Ordinary Differential Equations, Lotka-Volterra system, neural networks, flows.

1. Introduction

Models from the class of neural networks with an infinite impulse response such as recurrent neural networks, and more in particular Residual Neural Networks (ResNet), are the inspiration for the Neural Ordinary Differential Equations (NODE) network model [1]. In a residual network each layer can be defined as a finite transformation of the previous layer:

¹Corresponding Author: Argimiro Arratia, e-mail: argimiro@cs.upc.edu

$$h_{t+1} = h_t + g(h_t, \theta_t) \quad (1)$$

where h_t is the hidden state at layer t , g is a dimension preserving function and θ is a vector of parameters. These iterative updates can also be interpreted as an Euler discretization of a continuous transformation [3]. By augmenting the number of layers and taking smaller steps Δt , the ResNet dynamics expressed by (1) becomes in the limit an ordinary differential equation (ODE) specified by a neural network

$$\frac{dy(t)}{dt} = f(y(t), \theta(t), t) \quad (2)$$

Here f represents the Euler's discretization method to approximate a continuous function, and at each discrete time t , $y(t) = h_t$, $\theta(t) = \theta_t$ and $\Delta t f(h_t, \theta_t, t) = g(h_t, \theta_t)$. Thus, a neural ODE extends the traditional residual network model in that it has infinitely many layers at different point in time, and instead of fitting different weights and biases at each layer, it needs to fit a set of parameters θ that minimizes some cost function that depends on the initial values of $y(t)$, say $y(0) = h_0$, and the output of the system, say at some time T , $y(T) = h_T$ (w.l.o.g. we will consider $T = 1$). In other words, given the task of modeling a mapping $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, we are interested in optimizing the parameters θ such that the solution to the initial value problem

$$\frac{dy}{dt} = f(y, \theta, t), \quad y(0) = x \quad (3)$$

will accurately predict F at time $T = 1$. This value can be computed by an ODE solver, which evaluates the hidden unit dynamics of f wherever necessary from input layer $y(0) = h_0$ to output layer $y(1) = h_1$. There are many powerful and accurate optimizing algorithms, for instance, the family of algorithms based on the adjoint method [4].

Because we are interested in how the values of $y(1)$ change given the initial values $y(0)$, we study the *flow* of the NODEs. The NODE approach promises to yields benefits in terms of memory management as well as harnessing the theoretical wealth of minimization techniques, and stability results, in dynamical systems.

A first step to explore this approach is to study the universality of the NODEs in terms of their flows being able to approximate any function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Results have been obtained that give topological and analytical conditions on collections of functions such that their flows are universal for different measures of approximations [6]. Most of these approaches, however, study NODEs where the function $f(y, \theta, t)$ itself is a neural network, e.g. [5]. We note that these characterizations do not generate in themselves interesting families of SODEs.

We are interested in a more direct approach. We want to identify first families \mathcal{G} of systems of differential equations that are universal, in the sense that they encompass most of the systems of differential equations that appear in practice. Then we want to find a family \mathcal{F} of functions f for which the flows of the differential equation in (1) is exactly the family \mathcal{F} . Thirdly, we want to explore the process of optimizing the parameters of a differential equation (1) for functions f in \mathcal{F} .

In the following sections, we first present our general scheme for defining flows of universal SODEs and then the Lotka-Volterra system as our first candidate of universal SODE for building Neural ODE.

2. Computing the flow

Definition 2.1 Let $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a time independent vector field and let $\mathbf{y}(t)$ be the solution of the Initial Value Problem (IVP)

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}(t), \theta) \text{ with } \mathbf{y}(0) = \mathbf{x}.$$

Then the function

$$\varphi : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n,$$

defined as $\varphi(\mathbf{x}, t) = \mathbf{y}(t)$ is called the flow of the vector field \mathbf{f} .

We can define the optimization process the neural ODE appearing in (3) in terms of the flow as follows:

Given the IVP in (3), the optimization process (by the gradient method based on the adjoint approach) obtains the values of the parameters θ that minimize a total error (or cost function) constructed from comparing a collection of training data

$$\hat{T} = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i \in I}$$

with a collection of calculated outputs of the IVP in (3):

$$T = \{(\mathbf{x}_i, \varphi(\mathbf{x}_i, 1))\}_{i \in I}.$$

Once the optimal values of θ are obtained, the solution to the optimization problem is given by the function $\varphi(\mathbf{x}, 1)$. More specifically, for any input \mathbf{x} , the value of $\varphi(\mathbf{x}, 1)$ is the output of neural SODE solution of the IVP (3) at $t = 1$, with the optimal parameters and with initial value $\mathbf{y}(0) = \mathbf{x}$.

Observe that for the optimization process and for the generation of the approximation functions, the key role is played by the flow value $\varphi(\mathbf{x}, 1)$ and not by the solution of the IVP in (3) in the interval $(0, 1)$ of t .

In order for the Neural SODEs approach to be useful, one needs to ensure that the flows of the SODEs involved are **universal**, in the sense that there exists a family \mathcal{S} of SODEs such that any function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ can be approximated by the flow of an SODE $S \in \mathcal{S}$. Here the definition of “approximate” can take many different meanings, as discussed elsewhere, but essentially means that for any F , there exists an $S \in \mathcal{S}$ with associated flow φ_S such that for any $\mathbf{x} \in \mathbb{R}^n$, $F(\mathbf{x})$ is close to $\varphi_S(\mathbf{x}, 1)$.

Here is how we propose to find a collection \mathcal{S} of SODEs that is universal. First, we simplify the situation above as follows. Observe that for every natural n there exists (computationally) easy bijections \mathbb{R} into \mathbb{R}^n . Let us call one such (computable) bijection \mathbf{h} . Thus we can recast approximating a function $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ into approximating an equivalent function $\mathbf{F}^* : \mathbb{R} \rightarrow \mathbb{R}^n$ defined as $\mathbf{F}^*(t) = \mathbf{F}(\mathbf{h}(t))$. Thus we can recast the following concepts:

1. The IVP (3) is now: $\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}, \theta)$ with $\mathbf{y}(0) = (x, x, \dots, x)$.

2. The training data set $\hat{T} = \{(x_i, \hat{y}_i)\}_{i \in I}$ is now of the form $\hat{T}_1 = \{(x_i, \hat{y}_i)\}_{i \in I}$.
3. The set of inputs-outputs of the IVP is now $T_1 = \{(x_i, \varphi(x_i, 1))\}_{i \in I}$, where $\varphi(x, 1)$ is the solution of the IVP with initial condition $y(0) = (x, x, \dots, x)$, at $t = 1$.

Note that $\varphi(x, 1)$ is a function from \mathbb{R} to \mathbb{R}^n . Hence our optimization process for neural SODE becomes the following.

Definition 2.2 (The optimization process for a Neural SODEs) *Given an IVP:*

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}, \theta) \text{ with } \mathbf{y}(0) = (x, x, \dots, x) \quad (4)$$

The optimization process (by the gradient method based on the adjoint approach) obtains the values of the parameters θ that minimize a total error (or cost function) constructed from comparing a collection of training data

$$\hat{T} = \{(x_i, \hat{y}_i)\}_{i \in I}$$

with a collection of calculated outputs of the IVP in (4):

$$T = \{(x_i, \varphi(x_i, 1))\}_{i \in I}.$$

Once the optimal values of θ are obtained, the solution to the optimization problem is given by the function $\varphi(x, 1)$. More specifically, for any input x , the value of $\varphi(x, 1)$ is the output of our optimized Neural SODE.

3. The Lotka-Volterra system

Now we consider the Lotka-Volterra n -dimensional SODE. We begin by observing, as done elsewhere, that there is heuristic evidence that the families of n -dimensional Lotka-Volterra are universal in the sense that most of the types of systems of differential equations used in practice can be reduced to SODE's that are n -dimensional Lotka-Volterra [2]. Thus, a vast collection of functions F , the solutions of a very general collection of SODEs, can be seen as solutions of a Lotka Volterra n -dimensional SODE.

We propose then to study the flows $\varphi(x, t)$ that are associated with n -dimensional Lotka-Volterra SODEs. More specifically, given an n -dimensional Lotka Volterra Initial Value Problem:

$$LV(n, \theta, \mathbf{z}_0) : \frac{d\mathbf{z}}{dx} = \mathbf{L}(\mathbf{z}, \theta) \text{ with } \mathbf{z}(0) = \mathbf{z}_0, \quad (5)$$

where the n -dimensional Lotka-Volterra equations $LV(n, \theta, \mathbf{z}_0)$ with parameters $\theta = (\lambda_1, \dots, \lambda_n; A_{ij} : 1 \leq i \leq n, 1 \leq j \leq m)$ are:

$$z'_i = \lambda_i z_i + z_i \sum_{j=1}^m A_{ij} z_j, \quad i = 1, \dots, n$$

we will associate to it a (trivial) SODE with flow $\varphi_{n,\theta,z_0}(t,x)$ such that $\varphi_{n,\theta,z_0}(x,1) = \mathbf{z}(x)$.

Since the Lotka Volterra SODE's capture all the useful SODEs, we can hope that the functions $\varphi_{n,\theta,z_0}(\cdot, 1)$ will capture all solutions of the useful SODEs.

Our proposed process to use Neural SODE's on a training problem defined by a collection of points $\hat{T} = \{(x_i, \hat{\mathbf{y}}_i)\}_{i \in I}$, with, for every i , $\hat{\mathbf{y}}_i \in \mathbb{R}^n$ is explained in the following section.

3.1. Proposed optimization Process for Neural SODEs

We consider the Lotka-Volterra n -dimensional IVP defined in (5):

$$LV(n, \theta, \mathbf{z}_0) : \frac{d\mathbf{z}}{dx} = \mathbf{L}(\mathbf{z}, \theta) \text{ with } \mathbf{z}(0) = \mathbf{z}_0$$

We define an associated IVP for a SODE in the independent variable t and with parameters θ, t as follows

$$\frac{d\mathbf{y}}{dt} = \begin{bmatrix} -x \\ -x \\ \dots \\ -x \end{bmatrix} + \mathbf{z}(x) \text{ with } \mathbf{y}(0) = \begin{bmatrix} x \\ x \\ \dots \\ x \end{bmatrix}, \quad (6)$$

where $\mathbf{z}(x)$ is the solution of the $LV(n, \theta, \mathbf{z}_0)$ defined in (5), evaluated at x . Since the solution of the IVP (6) is simply

$$\mathbf{y}(t, x) = (1-t) \begin{bmatrix} x \\ x \\ \dots \\ x \end{bmatrix} + (t) \mathbf{z}(x).$$

We see that $\mathbf{y}(1, x) = \mathbf{z}(x)$. Since $\varphi(x, 1)$, the flow of the IVP (6) at $t = 1$, is $\mathbf{y}(1, x)$, we have that $\varphi(x, 1) = \mathbf{z}(x)$.

We apply the optimization process (by the gradient method based on the adjoint approach) to $LV(n, \theta, \mathbf{z}_0)$ to obtain the values of the parameters (θ, \mathbf{z}_0) that minimize a total error (or cost function) constructed from comparing the collection of training data

$$\hat{T}_1 = \{(x_i, \hat{\mathbf{y}}_i)\} i \in I$$

with a collection of calculated outputs of the $LV(n, \theta, \mathbf{z}_0)$, and defined as

$$T = \{(x_i, \varphi(x_i, 1))\}_{i \in I} = \{(x_i, \mathbf{z}(x_i))\}_{i \in I},$$

(since $\varphi(x, 1) = \mathbf{z}(x)$).

Once the optimal values of θ, \mathbf{z}_0 are obtained, the solution to the optimization problem is given by the function $\mathbf{z}(x)$. More specifically, for any input x , the value of $\varphi(x, 1) = \mathbf{z}(x)$ is the output of our approximation function.

We remark that in this approach, we are optimizing the parameters θ, \mathbf{z}_0 of the initial value problem $LV(n, \theta, \mathbf{z}_0)$ using the adjoint method, as in the classical Neural SODE approach. The advantages of this approach are: the process is easily adapted to other universal families of SODEs, such as Riccati equations; the process is easily scalable to any dimension n ; as with the canonical Neural SODE approach, the use of memory may be limited in this approach.

4. Experiments

4.1. Reproducing the evolution of Lotka-Volterra populations

We initialized a Neural ODE whose differential equation corresponds to an n -dimensional Lotka-Volterra system with random parameters $\theta \in [-0.05, 0.05]^{n(n+1)}$ and initial conditions $\mathbf{z}_0 \in [0, 1]^n$. The goal was to find such optimal values that minimize a loss function given by

$$\text{Loss}(\hat{\mathbf{Y}}, \mathbf{Y}) = \sum_i |\hat{\mathbf{y}}_i - \mathbf{y}_i|^2. \quad (7)$$

The first experiment was performed on an easy data set produced by solving a Lotka-Volterra system of equations. Populations following n -dimensional Lotka-Volterra systems for $n = 2, 5, 10$ were generated by (randomly) choosing values of parameters $\hat{\theta} \in [-0.25, 0.25]^{n(n+1)}$ and initial conditions $\hat{\mathbf{z}}_0 \in [0, 1]^n$ and letting the system evolve for a limited time span of $[0, 2]$. Population values $\hat{\mathbf{y}}_i$ at each time step x_i were recorded every 0.05 time units to generate the data sets $\hat{T}_n = \{(x_i, \hat{\mathbf{y}}_i)\}_i$. Random noise was added to $\hat{\mathbf{y}}_i$.

Results are shown in Figures 1 and 2. As we can see, our Neural ODE model parameters have been optimized correctly using the adjoint method.

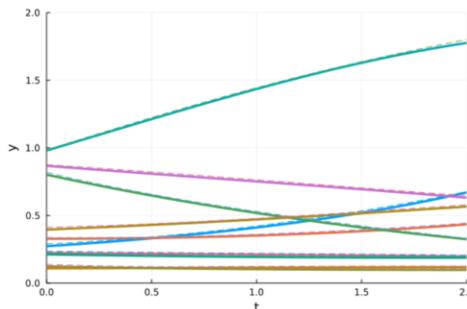


Figure 1. Plot of the estimated data (dashed lines) over the target data (thick lines, the noise added in the training set is not shown).

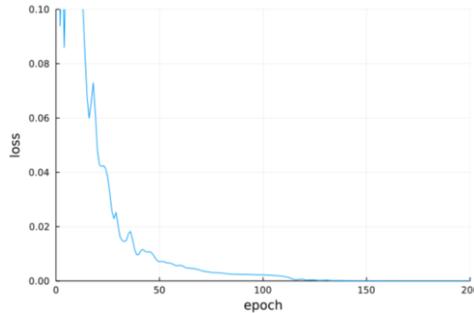


Figure 2. Loss of our Neural ODE model through the training phase.

4.2. Approximation of $1 + \sin(x)$

Now the goal is to approximate the function $1 + \sin(x)$ with a Neural ODE whose differential equation is a high dimensional ($n = 20$) Lotka-Volterra system. In this case, given that the target \hat{y}_i is 1-dimensional, the loss function is defined as to take into account only the first dimension of the system, while the other 19 may take whatever values are best.

Figure 3 shows that for a small interval $[0, \pi]$ our algorithm has been able to approximate the target function pretty well.

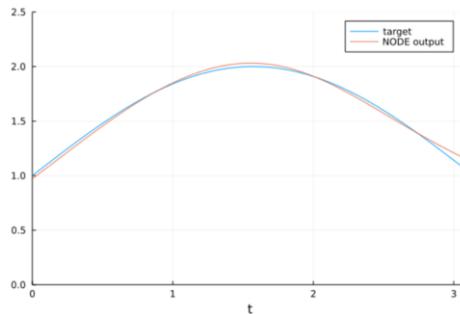


Figure 3. Plot of the 1st dimension of the Neural ODE output over the target $1 + \sin(x)$.

On the other hand, in Figure 4 we can see that even though the loss has not reached a minimum, numerical instabilities on the differential equation solvers complicate a further optimization.

References

- [1] Chen, R. T., Rubanova, Y., Bettencourt, J., Duvenaud, D. K. (2018). Neural ordinary differential equations. Advances in Neural Information Processing Systems, 31, pp. 6571-6583.
- [2] Hernández-Bermejo, B., Fairén, V. (1997). Lotka-Volterra representation of general nonlinear systems. Mathematical Biosciences, 140 (1), 1-32.
- [3] Lu, Y., Zhong, A., Li, Q., Dong, B. (2018). Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: International Conference on Machine Learning (pp. 3276-3285). PMLR.
- [4] Pontryagin, L. S. (1987). Mathematical theory of optimal processes. CRC press.

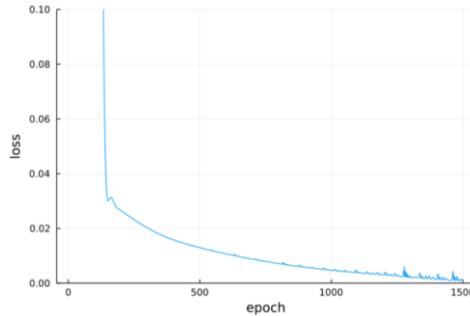


Figure 4. Loss of our Neural ODE model through the training phase.

- [5] Rackauckas, C., et al (2020). Universal differential equations for scientific machine learning. arXiv preprint arXiv:2001.04385.
- [6] Teshima, T., Tojo, K., Ikeda, M., Ishikawa, I., Oono, K. (2020). Universal approximation property of neural ordinary differential equations. arXiv preprint arXiv:2012.02414.

Garment Manipulation Dataset for Robot Learning by Demonstration Through a Virtual Reality Framework

Arnau BOIX-GRANELL ^{a,1}, Sergi FOIX ^a and Carme TORRAS ^a

^a*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain*

Abstract. Being able to teach complex capabilities, such as folding garments, to a bi-manual robot is a very challenging task, which is often tackled using learning from demonstration datasets. The few garment folding datasets available nowadays to the robotics research community are either gathered from human demonstrations or generated through simulation. The former have the huge problem of perceiving human action and transferring it to the dynamic control of the robot, while the latter requires coding human motion into the simulator in open loop, resulting in far-from-realistic movements. In this article, we present a reduced but very accurate dataset of human cloth folding demonstrations. The dataset is collected through a novel virtual reality (VR) framework we propose, based on Unity's 3D platform and the use of a HTC Vive Pro system. The framework is capable of simulating very realistic garments while allowing users to interact with them, in real time, through handheld controllers. By doing so, and thanks to the immersive experience, our framework gets rid of the gap between the human and robot perception-action loop, while simplifying data capture and resulting in more realistic samples.

Keywords. Garment manipulation, learning by demonstration, virtual reality framework, cloth folding dataset

1. Introduction

Non-rigid object manipulation has gained a lot of attention during the last decade since it has proven to be one of the big milestones to reach in the field of robotics in order to come closer to achieving full human-like capabilities. But the robotic manipulation of deformable objects is certainly not an easy task. There are two main difficulties that robots must face when manipulating a deformable object. On the one hand, there is the problem of fully estimating its state. Due to their ability to deform, non-rigid objects can take an infinite amount of configurations in space. Since fully observability is impossible to have in a real scenario, estimations must be made. Whereas rigid objects' pose can be easily estimated once a portion of its body is identified and located in 3D space, the correct deformable objects' state is nearly impossible to detect with just partial observability. On the other hand, there is the problem of gracefully manipulating a deformable object for fulfilling a task. Among others, factors such as the friction, elasticity and thickness of the

¹Corresponding Author: Arnau Boix-Granell, Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorenç i Artigas 4-6, 08028 Barcelona, Spain; E-mail: arnau.boix@estudiantat.upc.edu

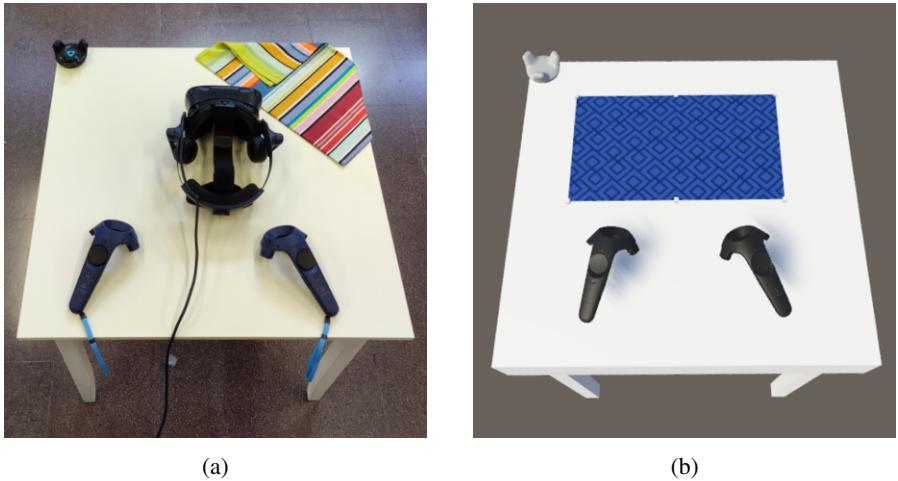


Figure 1. The manipulation is done on top of a real table that has its virtual version inside of the framework. To make sure that both of the objects are located in the same space (virtual and real worlds) we used HTC trackers for extrinsic calibration. (a) Real setup showing HTC's tracker (top left), headset (middle) and controllers (bottom). (b) Virtual setup showing HTC's tracker (top left), controllers (bottom) and simulated garment (center).

fabrics, the weight, size and shape of the garment, determine, not only the possible type of grasping, but also which actions can be taken and which ones not. Probably, due to these difficulties, there are not as many good datasets of deformable objects as there are of their rigid counterparts. This fact slows down the development of new artificial intelligence algorithms capable of understanding this type of objects, and therefore, creates a knowledge gap that this work pursues to fill.

One of the main challenges when trying to develop a garment-based dataset is whether to use real pieces of fabric or to use simulated ones. Currently, most of the available datasets are based on RGB-D images coming from real clothing data [1–8]. Despite the convenience of having real data, it is very hard to extract the ground truth information from garments and humans during a manipulation sequence. Moreover, data tend to have noise and multiple occlusions, and post-processing is always needed in order to have good estimated labeling. On the other hand, other approaches exploit the use of simulation environments to easily obtain fully observable ground truth data, although they must program the cloth manipulation behaviours with scripts. Therefore, this type of data lacks human-like demonstrations, losing the crucial manipulation dexterity contributions that would be provided by having the human perception into the loop. Imagine, for instance, the movement followed by a human hand previous to the prehension of a deformable object. That trajectory will, first, determine whether the grasping point will be successful or not and, second, which are going to be the next possible actions over that object in order to fulfill the assigned task. Recall that deformable objects may change their state after a manipulation and that, depending on that action, that change may be irreversible without adding extra manipulations.

In order to overcome those challenges, we propose a new approach that combines the use of simulated garments with human-based manipulation trajectories. Thanks to a virtual reality (VR) framework, humans can interact in real time with simulated pieces of

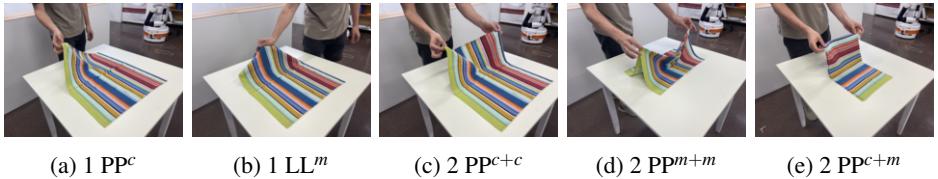


Figure 2. Classification of the different types of garment manipulations studied in this work: (a) One corner double point grasp (PP^c) with extrinsic planar contact (Π_e), (b) one middle edge double line grasp (LL^m) with (Π_e), (c) two corner double point grasp (PP^{c+c}) with (Π_e), (d) two middle edge double point grasp (PP^{m+m}), and (e) one corner, one middle edge double point grasp (PP^{c+m}) with (Π_e).

cloth (see Fig. 1a). In this work, we present a reduced but very accurate dataset of human cloth folding demonstrations.²

This article is structured as follows. Section 2 analyzes the related work in the literature. In Section 3 we present the different parts of our virtual reality set-up. Section 4 defines and explains our dataset storage format as an XML file. Section 5 introduces the different garments used in our experiments and the way we have classified them. Finally, Section 6 concludes this article.

2. Related Work

During the last decade, thanks to the creation of cloth manipulation datasets, a lot of progress in garment state detection and classification has been made. These datasets use either real or simulated fabrics in order to provide rich, and as accurate as possible, data reservoirs of cloth types, manipulation actions and garment states distribution.

2.1. Garment Datasets

In the context of garments, several attempts have been made to create various datasets. Some of those classify the garments by type [9–13], studying only static properties. Therefore, not useful when trying to understand manipulation processes. Others focus on the actions performed by a human when manipulating garments [14, 15]. Those works are mainly centred in studying the actions rather than the states of the piece of fabric, and for that reason, may not be as useful when trying to understand the evolution of garments between folding states. Others use RGB-D (or RGB) images to perceive the distribution of the garment [1–8]. These approaches have to estimate the occluded parts of the piece of fabric and, for that reason, might not be as helpful when high precision methods are required.

At the time of the writing, and despite the broad variety of approaches, the authors have no knowledge of any other studies that provide both the actions developed by a human while manipulating garments and, at the same time, the tracking the full evolution of the piece of fabric from an original state (before manipulation) to an ending state (after manipulation). As previously stated, our approach aims to fill this void.

²The dataset can be found in: <http://www.iri.upc.edu/groups/perception/clothingDataset/Data.rar>



Figure 3. Perception and Manipulation Lab's apartment mock-up.

2.2. Cloth State Manipulation

A problem encountered when starting to develop the dataset was to define a proper way to classify the different cloth states during a manipulation. As we already know, garments can have an infinite number of configurations, and, consequently, an infinite number of possible manipulations can be applied to them. In order to be able to plan a sequence of actions to take a deformable object from one state to another one, we must simplify the state-action representation. For that reason, some researchers have classified the types of manipulation based on both cloth and grasp type attributes, such as type of contact (single point, linear or planar), number of grippers used (single handed or bi-manual), or its final manipulation state. Some examples of these classifications can be found in [16].

For this work, we have classified the manipulations depending on the number of grippers used (one or two), the type of contact (single point P, linear L, or planar Π) and the part of the garment where the contact is made. We have used a classification method similar to the one showed in [17] (See Fig. 2).

Despite having chosen this method, due to the full observability properties and the recorded ground truth information of the data, any other type of garment manipulation classification could be applied. This has been one of the reasons for developing this framework, providing the community with a tool to test and compare different classification methods, given that we believe that the value for each classification method depends on the manipulation task performed.

3. Set-up

This work has been developed in the Perception and Manipulation Laboratory at the Institut de Robòtica i Informàtica Industrial (CSIC-UPC), using an *HTC Vive Pro* headset for creating an immersive VR experience thanks to the scenes created under the *Unity* framework (see Fig. 1). The Perception and Manipulation Lab hosts a life-scale mock-up of a fully-equipped apartment (see Fig. 3). Inside the apartment, researchers can study the interaction between robots and users in close-to-reality domestic environments.

3.1. HTC Vive Pro

As previously stated, we wanted to develop a framework that not only allows the visualization of garments, but also allows to create realistic garment manipulations. In order to do so, we believed that virtual reality could create the desired interactive experience. With that objective, we used an *HTC Vive Pro*, a virtual reality headset developed by *HTC Corporation* in collaboration with *Valve Corporation*. This type of devices are famous for offering the possibility of entering into an immersive experience in which the user is able to interact into Virtual Reality (VR), Mixed Reality (MR) or Augmented Reality (AR) worlds.

Despite of its main use, *HTC* headsets are also used by developers in other fields, for example, several research teams are developing AR applications to enhance the learning of manual assemblies [18]. Besides the headset, the *HTC Corporation* also offers some accessories that help making a more immersive experience. The two devices used in this project are the *HTC Tracker* and the *HTC Controller* (see Fig. 1). The tracker eases the connection between the real and the virtual world, making it possible to connect virtual objects with its real counterpart (as long as it has the tracker attached). The controller not only sends its real position to the virtual world but it can also send some basic information using its integrated buttons. More precisely, the *HTC Vive's* controller offers a total of three different buttons, one pressure-sensitive trigger and a trackpad.

The *HTC* hardware can easily be connected to *Unity* downloading *SteamVR*'s application and its asset (downloadable from *Steam*'s store and *Unity*'s asset store, respectively). The asset implements basic prefabs that allow the creation of VR experiences where all of *HTC* components can be used.

3.2. Unity

For the development of the framework, we decided to use the *Unity* engine. *Unity* is a cross-platform game engine developed by *Unity Technologies* [19]. *Unity*'s engine is mainly used for game developing due to its versatile and easygoing interface. Despite of that, it is also used for several engineering and AI applications. For example, the implementation of intelligent agents capable of overcoming obstacles or solving basic games such as mazes or arcade-like games [20–23].

In this work, *Unity* is used to build a framework where the information coming from the *HTC Vive Pro* system is displayed in a 3D environment. Moreover, the game engine will also work as a data reading and processing tool. Out of the possible simulators that could have been used to develop this work *Unity* was chosen for having fast simulation and providing a flexible control [24]. On top of that, the game engine was also used because of its user-friendliness, allowing future research groups interested in this framework to reproduce it or to apply their own changes to the simulations, if desired.

3.2.1. Obi Cloth Unity asset

Once a simulation engine was chosen, the next step was to study how to simulate the garments within the 3D environment. After some research into the different asset extensions for cloth simulation within *Unity*, we discovered a dedicated collection of particle-based physics plugins for deformable objects, such as cloth, fluids, ropes and soft-bodies, called *Obi* [25]. Every *Obi* object is made by a set of particles that can interact with each

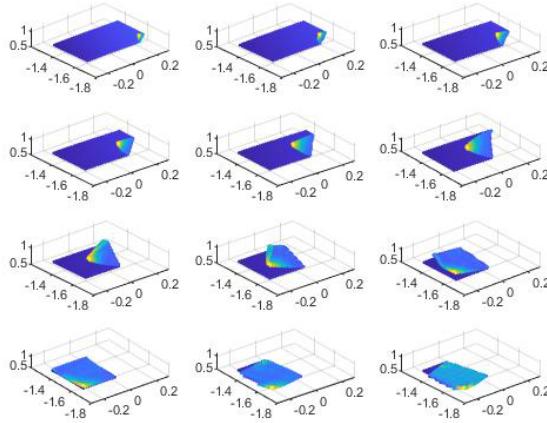


Figure 4. Sequence of point clouds obtained during a $PP^c + \Pi_e$ manipulation, with a z-axis heat map.

other, and affect or be affected by other objects within the scene. Moreover, particles can be constrained to have a customized behaviour. Compared to the other physics systems available in *Unity*'s assets store, Obi Cloth asset goes one step further by allowing much more constraints per cloth and by setting each particle's restriction separately.

4. Methodology

This section provides a brief explanation on how the dataset has been collected. Each manipulation is stored in a XML document with three main fields: Name, Mesh and Frames. The Name field corresponds to a string representing the name's experiment that has been performed. The Mesh field indicates the index of all the vertices that create a mesh element. Finally, the Frames field stores the evolution of the data at each timestamp.

For this dataset, we wanted to keep track of all the elements involved in a cloth manipulation task. In our current experiments, four elements were completely tracked. The first one, the garment per se. The dataset collects the coordinates of each particle of the fabrics, saving it under the tag name of *vertices*, inside of the geometry field within each frame (see Fig. 4). In order to easily export each mesh frame, data has been recorded maintaining *Ogre*'s mesh XML data-structure [26]. Secondly, we wanted to keep track of each of the *HTC* controllers used for manipulating the garments. In the case of a bi-manual operation, the tag names for each controller are *ControllerRight* and *ControllerLeft*. For each controller we store its pose components (position and rotation), a variable telling whether a grasping point is being held and a variable tracking the state of the trigger. This value was added thinking about future upgrades where changing the pressure over the surface of the grasped objects could be necessary for carrying out tasks such as edge tracing. Thirdly, it is also important to keep track of the position and rotation of the grasping points. Besides from that, and similarly to the controllers, we added a variable that indicates whether the right or left controller is holding the object or not. That third object type can be found under the tag of *GripPoint[i]*, where *i* is an integer value

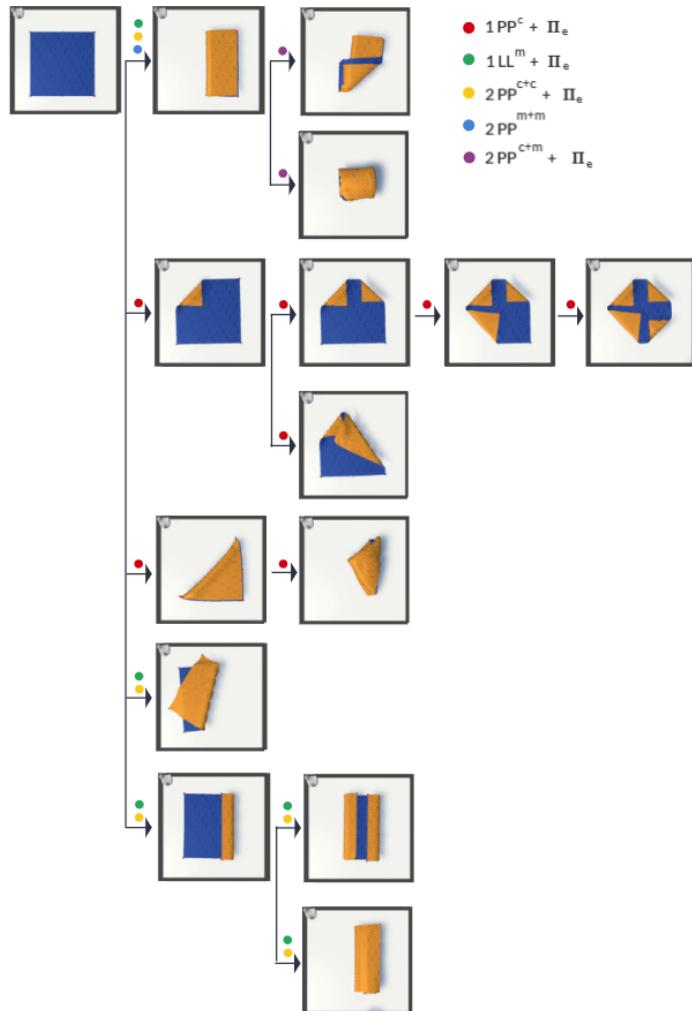


Figure 5. Graph of possible states for the napkin garment classified by manipulation using [17] method. The coloured dots indicate the different types of manipulation that can be performed in order pass from the previous state to the next one.

between one and the number of total simulated grasping points. Finally, the last object added to the dataset is the simulated table. From this object, the position and rotation are recorded under the name of Table.

All the experiments within the dataset have been conducted by a human using the HTC controllers as grippers. All the stated variables are recorded in XML files at 10 Hz. Finally, the dataset comes with two .txt files explaining both the format template of the XML documents and the data subsets distribution format.

5. Experiments

For a better versatility of the collected data, the conducted experiments have been divided into states. Each experiment starts in one described state and ends into another. With that methodology, if data of new experiments were required, only the new states would have to be recorded, given that the processes that follow the same sequence of actions can be reused.

A total of three different garments were used in these experiments, the properties of which can be seen in Table 1. These garments have been extracted from the household cloth object set studied in [27].

Table 1. Types of garment used in the experiments.

Name	Size [m]	Weight [kg]
Small Towel	0.3 x 0.5	0.08
Napkin	0.5 x 0.5	0.05
Tablecloth	0.90 x 1.30	0.188

In order to keep the dataset as brief and as rich as possible, we tried to just perform the most representative garment manipulations that are equivalent to all the studied garments. We use both single handed and bi-manual interactions, and we used them over different combinations of point, line and plane contact types. Due to the data format, it is easy to filter the manipulations by contact or interaction types with the objective of applying learning algorithms. Fig. 5 shows the complete sequence of states that have been studied for the *Napkin* case. As shown in the graph, some states can be achieved by performing different types of manipulation. For this garment, a total of nineteen manipulation sequences have been performed, with three repetitions each. These sequences correspond to all of the possible combinations of manipulations that start with the top-left state of Fig. 5 and end with one of the states on the right of the image. Whereas the *Small Towel* garment shares nearly the same state transition diagram, the *Tablecloth* garment is far too big for carrying on the examples within the state transition diagram. Despite that, we have included into our dataset a special case where the *Tablecloth* garment is hanging from a bar and has to set on the table thanks to a bi-manual manipulation and by taking advantage of the dynamics of the fabrics (see Fig. 6).

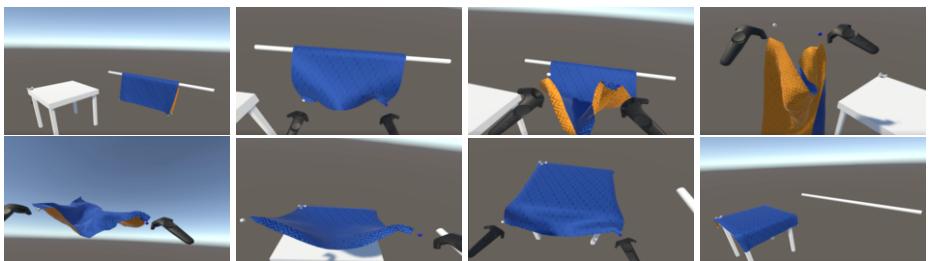


Figure 6. Manipulation of Tablecloth: From initial hanging position (top-left) to set on the table (bottom-right). The images in-between show frames from the two corner double point grasp manipulation performed on the Tablecloth garment.

6. Conclusions

In this work, we presented a *Unity* virtual reality framework to perform garment manipulation experiments. The approach used on the development differs from others in that we not only perform a full-mesh tracking but we also keep track of the position, rotation and interactions of other key features of the manipulation (like grippers or grasping points). Moreover, the implementation of the virtual reality allows the creation of an immersive experience that gets rid of the gap between the human and robot perception-action loop.

Later, we use the developed framework to create a rectangular garment manipulation dataset which is divided in states to allow a more versatile study. This new dataset aims to help the garment manipulation AI community by providing more realistic human-like garment manipulation data, which can be used in learning-from-demonstration approaches.

As a future work, we are planning on implementing a way to keep track of the states of the garments, by providing data such as how many corners are folded or if part of the garment is on top of another.

Acknowledgments

This work was developed in the context of the project CLOTHILDE (“CLOTH manipulation Learning from DEMonstrations”) which has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 741930) and is also supported by the BURG project PCI2019-103447 funded by MCIN/ AEI /10.13039/501100011033 and by the “European Union”.

References

- [1] Kimitoshi Yamazaki and Masayuki Inaba. Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2710–2717. IEEE, 2013.
- [2] Ioannis Mariolakis and Sotiris Malassiotis. Matching folded garments to unfolded templates using robust shape analysis techniques. In *International Conference on Computer Analysis of Images and Patterns*, pages 193–200. Springer, 2013.
- [3] Andreas Doumanoglou, Andreas Kargakos, Tae-Kyun Kim, and Sotiris Malassiotis. Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 987–993. IEEE, 2014.
- [4] Gerardo Aragon-Camarasa, Susanne B Oehler, Yuan Liu, Sun Li, Paul Cockshott, and J Paul Siebert. Glasgow’s stereo image database of garments. *arXiv preprint arXiv:1311.7295*, 2013.
- [5] Bryan Willimon, Ian Walker, and Stan Birchfield. A new approach to clothing classification using mid-level layers. In *2013 IEEE International Conference on Robotics and Automation*, pages 4271–4278. IEEE, 2013.
- [6] Georgios Tzelepis, Eren Erdal Aksoy, Júlia Borràs, and Guillem Alenyà. Semantic state estimation in cloth manipulation tasks. *arXiv preprint arXiv:2203.11647*, 2022.
- [7] Arnau Ramisa, Guillem Alenyà, Francesc Moreno-Noguer, and Carme Torras. Learning rgb-d descriptors of garment parts for informed robot grasping. *Engineering Applications of Artificial Intelligence*, 35:246–258, 2014.
- [8] Enric Corona, Guillem Alenyà, Antonio Gabas, and Carme Torras. Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition*, 74:629–641, 2018.

- [9] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [10] Thomas Ziegler, Judith Butepage, Michael C Welle, Anastasiia Varava, Tonci Novkovic, and Danica Kragic. Fashion landmark detection and category classification for robotics. In *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pages 81–88. IEEE, 2020.
- [11] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *European Conference on Computer Vision*, pages 512–530. Springer, 2020.
- [12] Li Sun, Gerardo Aragon-Camarasa, Simon Rogers, Rustam Stolkin, and J Paul Siebert. Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6699–6706. IEEE, 2017.
- [13] Li Sun, Simon Rogers, Gerardo Aragon-Camarasa, and J Paul Siebert. Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 2464–2470. IEEE, 2016.
- [14] John Schulman, Alex Lee, Jonathan Ho, and Pieter Abbeel. Tracking deformable objects with point clouds. In *2013 IEEE International Conference on Robotics and Automation*, pages 1130–1137. IEEE, 2013.
- [15] Andreas Verleysen, Matthijs Biondina, and Francis Wyffels. Video dataset of human demonstrations of folding clothing for robotic folding. *The International Journal of Robotics Research*, 39(9):1031–1036, 2020.
- [16] Júlia Borràs, Guillem Alenyà, and Carme Torras. A grasping-centered analysis for cloth manipulation. *IEEE Transactions on Robotics*, 36(3):924–936, 2020.
- [17] J. Borràs I. Garcia-Camacho and G. Alenyà. Knowledge representation to enable high-level planning in cloth manipulation tasks. *ICAPS 2022 Workshop on Knowledge Engineering for Planning and Scheduling*, 2022.
- [18] Yun Zhou, Shangpeng Ji, Tao Xu, and Zi Wang. Promoting knowledge construction: a model for using virtual reality interaction to enhance learning. *Procedia computer science*, 130:239–246, 2018.
- [19] Unity Technologies. [online] unity documentation. <https://docs.unity3d.com/Manual/index.html>, 2021.
- [20] Tom Ward, Andrew Bolt, Nik Hemmings, Simon Carter, Manuel Sanchez, Ricardo Barreira, Seb Noury, Keith Anderson, Jay Lemmon, Jonathan Coe, et al. Using unity to help solve intelligence. *arXiv preprint arXiv:2011.09294*, 2020.
- [21] A Juliani, VP Berges, E Vckay, Y Gao, H Henry, M Mattar, and D Lange. Unity: A general platform for intelligent agents. arxiv 2018. *arXiv preprint arXiv:1809.02627*.
- [22] Lucas Alberto E Pineda Metz. *An evaluation of unity ML-Agents toolkit for learning boss strategies*. PhD thesis, 2020.
- [23] Maryam Honari. Unity-technologies ml-agents. <https://github.com/Unity-Technologies/ml-agents>, 2013.
- [24] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018.
- [25] Virtual Methods Studio. [online] obi documentation. <http://obi.virtualmethodstudio.com/>, 2022.
- [26] Steve Streeting and Rojtberg Pavel. Ogre mesh xml. <https://github.com/OGRECommunity/ogre/blob/master/Tools/XMLConverter/docs/ogremeshxml.dtd>, 2018.
- [27] Irene Garcia-Camacho, Júlia Borràs, Berk Calli, Adam Norton, and Guillem Alenyà. Household cloth object set: Fostering benchmarking in deformable object manipulation. *IEEE Robotics and Automation Letters*, 7(3):5866–5873, 2022.

Long Short-Term Memory to Predict 3D Amino Acids Positions in GPCR Molecular Dynamics

Juan Manuel LÓPEZ-CORREA ^{a,1}, Caroline KÖNIG ^{a,b} and Alfredo VELLIDO ^{a,b}

^aComputer Science Dept., Univ. Politècnica de Catalunya - UPC BarcelonaTech,
08034, Barcelona, Spain

^bIntelligent Data Science and Artificial Intelligence (IDEAI-UPC) Research Center

Abstract. G-Protein Coupled Receptors (GPCRs) are a big family of eukaryotic cell transmembrane proteins, responsible for numerous biological processes. From a practical viewpoint around 34% of the drugs approved by the US Food and Drug Administration target these receptors. They can be analyzed from their simulated molecular dynamics, including the prediction of their behavior in the presence of drugs. In this paper, the capability of Long Short-Term Memory Networks (LSTMs) are evaluated to learn and predict the molecular dynamic trajectories of a receptor. Several models were trained with the 3D position of the amino acids of the receptor considering different transformations on the position of the amino acid, such as their centers of mass, the geometric centers and the position of the α -carbon for each amino acid. The error of the prediction of the position was evaluated by the mean average error (MAE) and root-mean-square deviation (RMSD). The LSTM models show a robust performance, with results comparable to the state-of-the-art in non-dynamic 3D predictions. The best MAE and RMSD values were found for the mass center of the amino acids with 0.078 Å and 0.156 Å respectively. This work shows the potential of LSTM to predict the molecular dynamics of GPRCs.

Keywords. G-Protein Coupled Receptors, LSTM, Molecular Dynamics

1. Introduction

G-Protein Coupled Receptors (GPCRs) are a big family of eukaryotic cell transmembrane proteins. They are abundant cell surface receptors accounting for 4% (800) of all human genes [1] and responsible for numerous biological processes. This is the result of their ability to transmit extracellular signals, which makes them relevant for pharmacology and the research about drugs targeting these receptors. Around the 34 % of the drugs approved by the US Food and Drug Administration [1]. This has led, over the last decade, to active research in the field of proteomics.

¹Corresponding Author: Juan Manuel LÓPEZ-CORREA, Univ. Politècnica de Catalunya. Barcelona Tech, 08034, Barcelona, Spain juan.manuel.lopez.correa@upc.edu.

The functionality of a protein depends widely on its 3-D structure, which determines its ability for certain ligand binding. However, the 3-D structure of human GPCRs is not fully determined yet [2]. As an alternative, when the information about the 3-D structure is not available, the investigation of the functionality of a protein can be achieved through the analysis of its amino acid sequence, which is known and available in several public curated databases[3]. Computing biotechnology, X-ray crystallography and cryo-electron microscopy over the last year has evolved exponentially, yielding 3-D models of many proteins. Such structures are publicly available at repositories, such as at the Protein Data Bank (PDB) [4], but provide only a static view of the receptor's state. Molecular dynamics (MD) simulations are an interesting technology to explore dynamically the conformational landscape of receptors under the presence of different drugs. MD simulations take the 3D models as starting point for the computer assisted simulation and explore the molecular dynamics at the simulated environment [5]. Specific repositories of simulated MDs are available, such as the GPCRMD for MD simulations of GPCRs [6].

For the study of the temporal evolution of molecular dynamics with machine learning (ML) techniques, Recurrent Neural Networks (RNN) are used due to their capability for modeling temporal sequences. RNNs have shown success at applications such as human language modeling [7]. In recent years, a particular RNN, Long Short-Term Memory (LSTM)[8], has been successfully applied to machine translation[9][10], speech recognition [11], sequence learning [12] and weather forecasting [13]. LSTMs solve a limitation of the RNN architecture, the inability to learn information originated from far past in time. LSTMs overcome that limitation by their ability to accumulate information for a long period of time and allowing the network to dynamically learn to forget old aspects of information. Recently, LSTMs have been used to mimic trajectories produced by simulations[14], achieving accurate predictions about a short time into the future. LSTM and their variants have shown great potential in sequence processing [15], and there are several studies where they were applied for the analysis of trajectories from simulation systems [16]. In [17] the temporal evolution of chemical/biophysical trajectories is predicted by LSTMs. Mohamma *et al.* (2019) [18] used LSTMs to predict interactions between atoms over the simulation time to temporary correlations among them. Kadupitiya *et al.* (2020) [19] incorporate LSTM into the numerical integrator that solves Newton's equations in molecular dynamics simulations [19]. Other authors applies LSTM directly onto the low dimensional molecular trajectories and predicts the rare events in the sequential data [20]. Liang *et al.* (2019)[21] apply LSTMs to trajectories forecasting of spike glycoproteins (S-protein) on the SARS-CoV-2 dynamics. Ludwig *et al.* (2022) have worked with special LSTM, bi-directional, to increase the 3D spacial resolution of MD trajectories within a post-processing step. However, none of them have reported results to predict the amino acid's position in the molecular dynamics of GPCRs. Yadav *et al.* (2020) [22] developed 3 machine learning approaches to predict the conformation state of GPCR proteins obtaining similar errors (MAE: 0.0715 - 0.0897 Å and RMSD 0.1291 - 0.1449 Å) as it reported in this work but this developments was on no dynamics simulations.

In this work, the capability of LSTMs are evaluated to learn and predict the 3D positions of the amino acids of a GPCR receptor in molecular dynamic simulations. In a first experiment, the prediction of two types of LSTMs are compared, the unidirectional and bidirectional variant of LSTMs. In the second experiment, different representations

of the amino acid position and several variables of the LSTM are analyzed to find the combination of the parameters, which best predict the molecular trajectory. The experiments are carried out on a public available dataset of the molecular dynamic simulations of the β 2AR-rh1 GPCR receptor [23].

The remaining part of the articles is structured as follows. In section Materials the dataset under study is explained. Methods section describe the ML model, data preprocessing and the experimental setup. The Result section evaluates the quality of prediction of the models per experiment. Finally in the Discussion and Conclusion section the results and impact of the study are discussed.

2. Materials

2.1. MD simulations

In this work a dataset of the MD simulations of the β 2AR-rh1 GPCR receptor is analyzed. The β 2-adrenergic receptor (β 2AR) is implicated in type-2 diabetes, obesity, and asthma, and is a member of the class A, rhodopsin-like GPCRs (rh1) [24]. This simulations have been created by [23] at the Google's Exacycle cloud computing platform. The simulations under study in this work comprise 10.000 trajectories of the β 2AR-rh1 GPCR receptor with a full agonist. Each trajectory describes the 3D position of the receptor during 28 consecutive time-steps, which are referred to as frames in this study. The time elapsed between each frame are 500 picoseconds. The receptor has 282 amino acids for which the position is predicted during the different frames of the simulation in this work.

3. Methods

3.1. Long Short-Term Memory (LSTM)

A specific and extremely popular instance of RNNs are LSTM [8] neural networks, which show more flexibility and can be used for challenging tasks such as language modeling, machine translation, and weather forecasting [25,26,10]. In this paper, unidirectional LSTM (ULSTM) and bidirectional LSTM (BLSTM) are used to predict the trajectories of the MD simulations. ULSTMs work by processing data in the forward direction, while BLSTMs processes sequence data in both forward and backward directions with two separate hidden layers [27]. The bidirectional networks are often reported to yield better prediction results than unidirectional ones, such as at phoneme classification [28] or speech recognition [29], to number a few. Bidirectional LSTMs have not been used yet in molecular dynamic predictions of the GPCRs problem, based on a review of the literature [30,27,31,32,33].

3.2. Data Preprocessing

Data normalisation: The models are trained with normalized data. This process is done by applying a linear max–min normalization [34,35]. The normalized data pre-

dicted by the model can be transformed back to the original range of values to assess the quality of prediction in Angstrom (\AA) units.

Center of the amino acids: Each simulation is made up of 28 frames (step positions) and 282 amino acids. However, the original information of the MD simulation provides the position of the atoms, not of the amino acids. For this reason, three representations for the 3D amino acid position are calculated: a) *Geometric Center* (CG), b) *Center of Mass* (CM) and c) α -carbon (α C) conforming three derived datasets with the 3D positions (xyz) for each amino acid in each step (frame) of the simulated molecular trajectory.

3.3. Experimental Setup

In the following the variables and parameters used for the configuration of the experiments are explained:

nClones: As explained in the Materials section, the original dataset comprises 10.000 simulated trajectories of the β 2AR-rh1 receptor. Each of these trajectories is named a *nClones* in this work. Each trajectory is simulated under the presence of a full agonist.

nSteps-in : The training of the LSTM models is carried out with the information of short sequences of the 3D positions of the amino acids. The lengths of these sequences is named *nSteps-in* in this work. In one step each amino acid has three position values (x,y,z). In the experiments the models are trained with different values for the parameter *nSteps-in*.

nSteps-out: After the training of the LSTM models with different *nSteps-in*, the model can predict a sequence of next steps for the trajectory. The number of predicted steps is referred to as the parameter *nSteps-out*.

Model optimisation: Of the 10.000 *nClones* available at the original dataset, 1.000 were taken for model creation. This dataset was split in 5 folds with 200 *nClones* per fold. Four folds were used for training, conforming the train set and the remaining fold was used only for the evaluation of the quality of prediction of the model, conforming the test set. The model creation was carried out following a cross validation approach [9]. This means, the training was repeated 4 time per experiment. For each training repetition 3 folds of 4 train set folds were selected to train the model and the remaining fold (validation set) was used to test the predictions through *Mean Average Error* (MAE) [36]. This process was repeated for each fold of the train set. In this way, all the folds of the train set were tested without mixing the training and validation data. The training process generates 4 trained models and 4 testing results. The best model was chosen by the lowest MAE value. The best model was used to evaluate the predictive ability against new data never seen by the model,in this work the test set.

Outline of experiments: The first and second experiment were performed by two types of LSTMs - ULSTM and BLSTM. The first experiment was carried out comparing the predictions of the LSTM on the CG, CM, α C transformed dataset. The experiments aims to find out which representation of the amino acid center is best to predict the molecular dynamic sequences of the GPCR.

When a position sequence is shown to the LSTM, it needs to know the previous amino acid positions to predict the next positions. For this reason, the next experiment

investigates if the parameter *nSteps-in* has an impact on the quality of prediction. Values in the range of three, five and seven for the parameter *nSteps-in* per center of the amino acid were evaluated.

Finally, the third experiment was developed to know the capability of the ULSTM to predict different length of the sequence. For this experiments the predictions of 12 *nSteps-out* were evaluated.

The results in the following section are reported using two metrics calculated on the original range of values in Angstrom (\AA) unit:

1. Mean absolute error (MAE) [36] for each predicted value of the *x,y* and *z* position.
2. Root-mean-square deviation of amino acid positions (RMDS)[37] for each predicted value of *x,y* and *z* position.

4. Results

In this section the results of the experiments with LSTMs for the prediction of the sequences of the trajectories of the receptor are described. The ULSTM and BLSTM were trained to predict multivariate time series data on sample trajectories. Table 1 and Table 2 show the MAE (mean of the position "x", "y" and "z" for the amino acids), MAEx (mean of the position "x"), MAEy (mean of the position "y"), MAEZ (mean of the position "z") and RMSD by ULSTM and BLSTM respectively. In addition, in both tables the standard deviation [38] (std) for MAE and RMSD is indicated. In this experiment, to simplify the analysis the *nStep-out* was set to a value of 1. About the *nStep-in*, the results show the mean value from experiments carried out with the values 3,5,7 for the parameter *nSteps-in*.

Table 1. MAE predictions of the amino acid position by Geometric center, Mass center, α -carbon by ULSTM

Amino acid center	MAE	MAEx	MAEy	MAEZ	RMSD
Geometric	0.0850 ± 0.024	0.0793	0.0864	0.0894	0.1703 ± 0.040
Mass	0.0781 ± 0.016	0.0728	0.0793	0.0822	0.1561 ± 0.027
α -carbon	0.0792 ± 0.0219	0.0739	0.0809	0.0838	0.15934 ± 0.034

Table 2. MAE predictions of the amino acid position by Geometric center, Mass center, α -carbon by BLSTM

Amino acid center	MAE	MAEx	MAEy	MAEZ	RMSD
Geometric	0.0835 ± 0.024	0.0775	0.0845	0.0883	0.1673 ± 0.040
Mass	0.0806 ± 0.021	0.0753	0.0821	0.0844	0.1615 ± 0.025
α -carbon	0.0829 ± 0.023	0.0773	0.0842	0.0871	0.1661 ± 0.039

Figure 1 represents the results for the second experiment considering the *center of the mass* variable as amino acid representation. The MAE of the sequence prediction are shown considering the *nSteps-out* = 1 for ULSTM (rhombuses bar) and BLSTM (full gray bar) with three sequences lengths as *nSteps-in* = 3, 5, 7 values.

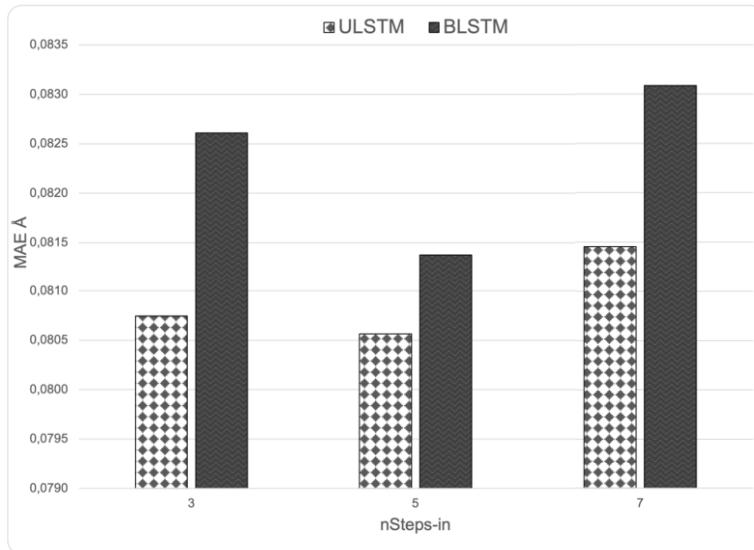


Figure 1. Mean Average Error(MAE) of the sequence prediction by ULSTM (full gray bar) and BLSTM (rhombuses bar) for three sequences lengths $nSteps-in = 3, 5, 7$ values. MAE in Å units

The objective of the third experiment is the evaluation of the forecasting capability of a ULSTM to predict long sequences. The MAE and RMSD behavior for the 12 $nSteps-out$ predictions is represented in the Table 3. In this experiment the parameters, which have yielded the best results in the previous experiments were used, i.e. center of mass representation, $nStepsIn$ of 5 and a ULSTM. In this case, the MAE predictions is discriminated in 3D coordinates (MAEx, MAEy, MAEz). As well also, the standard deviation (st) for mean 3D position MAE and RMSD is shown.

Table 3. Error prediction by MAE and RMSD for 12 length sequences ($nSteps-out$) with ULSTM.

step	MAE	MAEx	MAEy	MAEz	RMSD
1	0.0769 ± 0.0220	0.0713	0.0780	0.0814	0.1535 ± 0.0379
2	0.0814 ± 0.0224	0.0761	0.0823	0.0858	0.1626 ± 0.0381
3	0.0837 ± 0.0229	0.0784	0.0847	0.0880	0.1674 ± 0.0383
4	0.0842 ± 0.0231	0.0788	0.0855	0.0883	0.1684 ± 0.0388
5	0.0852 ± 0.0234	0.0797	0.0867	0.0892	0.1706 ± 0.0392
6	0.0854 ± 0.0234	0.0798	0.0869	0.0895	0.1712 ± 0.0390
7	0.0857 ± 0.0233	0.0797	0.0870	0.0903	0.1714 ± 0.0391
8	0.0862 ± 0.0236	0.0804	0.0870	0.0912	0.1725 ± 0.0396
9	0.0865 ± 0.0241	0.0810	0.0868	0.0919	0.1735 ± 0.0404
10	0.0864 ± 0.0241	0.0802	0.0869	0.0920	0.1730 ± 0.0412
11	0.08601 ± 0.02433	0.0793	0.0870	0.0916	0.1719 ± 0.0425
12	0.0861 ± 0.0246	0.0794	0.0872	0.0916	0.1722 ± 0.0433

5. Discussion

The first experiment focusing on the different transformations of the amino acid positions has shown that center of mass is the transformation that best predicts the molecular dynamics sequences of the GPRC yielding minimum MAE and std values for ULSTM and BLSTM (bold numbers in Table 1 and Table 2). However, the *geometric center* and *α-carbon center*, also, get quite good prediction results comparable with another works that work only with Static Molecular Structures [39,40]. The second experiment seeks to discover the best sequence length as *nSteps-in*. For both the ULSTM and the BLSTM the minimum MAE was obtained for the value of 5 for the input steps. This means 5 steps as the best length for the input information of the sequence for the model to best predict future trajectories. Regarding the capabilities of ULSTMs and BLSTMs, the ULSTM demonstrate a better performance in all experiments. Finally, the results of the third experiment reveal the increment of the error with the length of predicted sequence. With larger *nSteps-out* values, both the MAE and RMSD metric increase. In addition, the MAE in the plane "x" shows greater ability to predict more steps compared to the "y" and "z" plane. As well also, the plane z show the biggest error analysing the 3D coordinates.

6. Conclusions

GPCRs are family of receptors with great interest in pharmacology and molecular dynamics are a powerful tool to discover the conformational space and the behavior of the receptors. Due the large amount and complexity of data in MD simulations, machine learning approaches are a promising approach to discover relevant knowledge. This study has used a specific machine learning approach, namely LSTMs to study the ability to predict the movements of a receptor. This prediction is not trivial as the receptor comprises 282 amino acid, which yield a dataset of 846 data points. Furthermore, these datasets are in the context of a temporal sequence and methods taking into account the temporal evolution are needed. This study has demonstrated the potential of LSTMs to predict accurately molecular dynamics sequences of a GPRC receptor, specifically for β2AR-rh1. In addition, the study has provided insights about which are the best parameters regarding the representation of amino acid positions, the lengths of the input sequence and length of the predicted sequence. In particular, the *center of the mass* is the best representation of the 3D amino acid position for a complex receptor yielding the best results at the forecasting. Furthermore, the study has shown that the best length of input information are 5 steps. The prediction performance of ULSTM show slightly better results comparing with BLSTM, although both models achieved accurate results. Finally, the study also confirmed that the capability to predict long sequences decreases with the lengths of the forecasted sequence. These results are important for the configuration of other experiments on the analysis of MD data. As a future line of research the use of generative models is planned in order to artificially generate MD trajectories.

Acknowledgments

This work is funded by Spanish PID2019-104551RB-I00 research project and by the PhD. training program (PRE2020-092428) through the Ministry Science and Innovation of Spain.

References

- [1] Ismael Rodríguez-Espigares, Mariona Torrens-Fontanals, Johanna KS Tiemann, David Aranda-García, Juan Manuel Ramírez-Anguita, Tomasz Maciej Stepniewski, Nathalie Worp, Alejandro Varela-Rial, Adrián Morales-Pastor, Brian Medel-Lacruz, et al. Gpcrmd uncovers the dynamics of the 3d-gpcrome. *Nature Methods*, 17(8):777–787, 2020.
- [2] Vsevolod Katritch, Vadim Cherezov, and Raymond C Stevens. Structure-function of the g protein-coupled receptor superfamily. *Annual review of pharmacology and toxicology*, 53:531–556, 2013.
- [3] Caroline König, Raúl Cruz-Barbosa, René Alquézar, and Alfredo Vellido. Svm-based classification of class c gpcrs from alignment-free physicochemical transformations of their sequences. In *International Conference on Image Analysis and Processing*, pages 336–343. Springer, 2013.
- [4] Helen M Berman. The protein data bank: a historical perspective. *Acta Crystallographica Section A*, 64(1):88–95, 2008.
- [5] Naomi R Latorra, AJ Venkatakrishnan, and Ron O Dror. Gpcr dynamics: structures in motion. *Chemical reviews*, 117(1):139–155, 2017.
- [6] Ismael Rodríguez-Espigares, Mariona Torrens-Fontanals, Johanna KS Tiemann, David Aranda-García, Juan Manuel Ramírez-Anguita, Tomasz Maciej Stepniewski, Nathalie Worp, Alejandro Varela-Rial, Adrián Morales-Pastor, Brian Medel-Lacruz, et al. Gpcrmd uncovers the dynamics of the 3d-gpcrome. *Nature Methods*, 17(8):777–787, 2020.
- [7] R Rico-Martínez, IG Kevrekidis, MC Kube, and JL Hudson. Discrete-vs. continuous-time nonlinear signal processing: Attractors, transitions and parallel implementation issues. In *1993 American Control Conference*, pages 1475–1479. IEEE, 1993.
- [8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [9] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.
- [10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [11] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.
- [12] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
- [13] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- [14] Mohammad Javad Eslamibidgoli, Mehrdad Mokhtari, and Michael H Eikerling. Recurrent neural network-based model for accelerated trajectory analysis in aimd simulations. *arXiv preprint arXiv:1909.10124*, 2019.
- [15] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.
- [16] Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical review letters*, 120(2):024102, 2018.
- [17] Sun-Ting Tsai, En-Jui Kuo, and Pratyush Tiwary. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature communications*, 11(1):1–11, 2020.

- [18] Mohammad Javad Eslamibidgoli, Mehrdad Mokhtari, and Michael H Eikerling. Recurrent neural network-based model for accelerated trajectory analysis in aimd simulations. *arXiv preprint arXiv:1909.10124*, 2019.
- [19] JCS Kadupitiya, Geoffrey C Fox, and Vikram Jadhao. Deep learning based integrators for solving newton's equations with large timesteps. *arXiv preprint arXiv:2004.06493*, 2020.
- [20] Wensi Zeng, Siqin Cao, Xuhui Huang, and Yuan Yao. A note on learning rare events in molecular dynamics using lstm and transformer. *arXiv preprint arXiv:2107.06573*, 2021.
- [21] David Liang, Meichen Song, Ziyuan Niu, Peng Zhang, Miriam Rafailovich, and Yuefan Deng. Supervised machine learning approach to molecular dynamics forecast of sars-cov-2 spike glycoproteins at varying temperatures. *MRS advances*, 6(13):362–367, 2021.
- [22] Prakarsh Yadav, Parisa Mollaei, Zhonglin Cao, Yuyang Wang, and Amir Barati Farimani. Prediction of gpcr activity using machine learning. *Computational and Structural Biotechnology Journal*, 2022.
- [23] Joseph L Hellerstein, Kai J Kohlhoff, and David E Konerding. Science in the cloud: accelerating discovery in the 21st century. *IEEE Internet Computing*, 16(4):64–68, 2012.
- [24] Kai J Kohlhoff, Diwakar Shukla, Morgan Lawrenz, Gregory R Bowman, David E Konerding, Dan Belov, Russ B Altman, and Vijay S Pande. Cloud-based simulations on google exacycle reveal ligand modulation of gpcr activation pathways. *Nature chemistry*, 6(1):15–21, 2014.
- [25] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
- [26] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.
- [27] Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.
- [28] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.
- [29] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013.
- [30] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016.
- [31] Reza Paki, Esmaeil Nourani, and Davoud Farajzadeh. Classification of g protein-coupled receptors using attention mechanism. *Gene Reports*, 21:100882, 2020.
- [32] Xueliang Liu. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318*, 2017.
- [33] Xueliang Liu. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318*, 2017.
- [34] Ali Jahan and Kevin L Edwards. A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. *Materials & Design (1980-2015)*, 65:335–342, 2015.
- [35] MJ Asgharpour. Multiple criteria decision making. *Tehran: Tehran University*, 1998.
- [36] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development Discussions*, 7(1):1525–1534, 2014.
- [37] Karen Sargsyan, Cédric Graeffel, and Carmay Lim. How molecular size impacts rmsd applications in molecular dynamics simulations. *Journal of chemical theory and computation*, 13(4):1518–1524, 2017.
- [38] Dong Kyu Lee, Junyong In, and Sangseok Lee. Standard deviation and standard error of the mean. *Korean journal of anesthesiology*, 68(3):220, 2015.
- [39] Vitali Nesterov, Mario Wieser, and Volker Roth. 3dmolnet: a generative network for molecular structures. *arXiv preprint arXiv:2010.06477*, 2020.
- [40] Michael A Hanson and Raymond C Stevens. Discovery of new gpcr biology: one receptor structure at a time. *Structure*, 17(1):8–14, 2009.

This page intentionally left blank

Computer Vision

This page intentionally left blank

Towards Cross-Sites Generalization for Prostate MRI Segmentation to Unseen Data

Eddardaa BEN LOUSSAIEF^{a,1}, and Domenec PUIG^a

^aDepartment of Computer Engineering and Mathematics, Universitat Rovira i Virgili,
43007 Tarragona, Spain

Abstract. Learning a model from multi-source data is a challenging and topical learning problem. Thus, generalization capacity has been proposed to deal with the domain shift (i.e. various imaging vendors, modalities, and protocols) across domains. This paper tackles the out-of-distribution generalization for prostate segmentation in MRI imaging. We propose a simple approach based on the pretraining-finetuning scheme to boost the deep neural network's generalization to unseen data in prostate MRI segmentation. This paper introduces an objective loss that seeks to minimize cross-domain distribution by adapting Kullback–Leibler (KL) divergence. To manifest the effectiveness of our approach, we perform experiments on a multi-source public dataset for prostate MRI imaging collected from six vendors. As a result, the proposed model can yield promising cross-domains generalization capacity to unseen target domain.

Keywords. Prostate segmentation, MRI imaging, Domain generalization, Unseen target, Transfer learning.

1. Introduction

Data collection across multiple medical institutions is increasingly recommended to build an automated and accurate deep network for medical imaging analysis. Deep neural networks(DNNs) based medical imaging segmentation approaches have whitened advanced progress [1]. Much of the current literature on medical imaging segmentation plays particular attention to use transfer learning [4] and domain adaptation [2,3] to address this issue, however, all of them relies on a strong assumption that the images from the target domain are accessible (seen target) for model training. Instead, it has a broad prospect for clinical use to train a generalizable model that can learn from single or multiple related but distinct source domains in such way the model can be applied directly to any Out-Of-Distribution target domain.

We address this challenging concept to Domain generalization (DG) [5,6], in which during the training, no available prior knowledge from the test domain. DG has been introduced to deal with the domain shift problem. Under DG scope, domain-invariant rep-

¹Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain; E-mail: eddardaa.benloussaief@urv.cat.

resentation learning [9], Data manipulation [8] and meta-learning [6] approaches have achieved promising success for imaging segmentation. Transfer learning aims to train the model on source task and then improve the performance of the model where the source and the target have different tasks. It attempts to use pretraining-finetuning strategy while the target domain is accessed in training. Motivated by the strategy of transfer learning, we propose to combine the principle of pretraining-finetuning and domain alignment to tackle the generalization capacity to unseen domain for prostate MRI segmentation. Meanwhile, to ensure the distribution alignment, we adopt Kullback-Leibler (KL) divergence algorithm, that consist of matching the features from multiple source domains to a prior distribution.

2. Method

A basic deep learning model for medical imaging segmentation consist of feature extractor $F_\phi : X \rightarrow Z$ and a segmentor network $S_\theta : Z^2 \rightarrow Y$. We propose to use the paradigm of transfer learning that focus on acquiring knowledge from multiple sources to enhance the generalization in a related target. TL [4] is pivotal for developing strategies that address to domain shift across domains. It is beneficial for the reuse of trained deep model by using a fine-tuning strategy for the new task. We adopted a deep model to train each source domain separately, then we have selected the most efficient model based on dice score to reuse it as a pretrained model to train all the source domains together. Furthermore, we tried to fine-tune the pretrained model on the sources to obtain a generalizable model could perform well on the unseen target. We further propose to align the distribution by extracting shareable information i.e. feature representation. Motivated by [10], we propose to adopt an objective loss that is based on the Kullback-leibler(KL) divergence [12]. KL divergence aims to quantify the difference across probability distribution for a given seen domains. It seeks to match the latent features from seen sources to a pre-defined prior distribution. We adopt a Gaussian Distribution as a prior distribution $Z \sim \mathcal{N}(0, 1)$. We conduct a variational U-net [13] architecture to our work. We split our training into stages, first one consist of training and testing the model on each site separately, then select the best model. Next, we adopt the selected model from the previous stage as pretrained model by applying fine-tuning procedure to retrain the new segmentor on the multiple domains. We follow leave-one-domain-out strategy, we learn the model from the $K - 1$ domains and keep the one-left-out as unseen target domain. For example in our work, we adopt six sites, so for each experiment, we use five sites as multi-source domains and the remaining site to test the model's performance. We, furthermore, design a segmentation loss function, which mix between cross entropy loss, dice loss and KL divergence loss i.e. Gaussian loss for the distribution regularization. The segmentation loss should be optimized as following:

$$seg = \sum_{i,k} l_C(\hat{y}_i^k, y_i^k) + \sum_{i,k} l_D(1 - \left(\frac{(2\hat{y}_i^k y_i^k)}{(\hat{y}_i^k + y_i^k + \epsilon)} \right)) + \lambda l_{KL}$$

Where l_c and l_d denotes the cross entropy loss and the dice loss respectively. And λ is a smoothing weight to penalize the Gaussian loss.

²Z is the high-dimensional space for feature maps for segmentation task

3. Experiments

Dataset and evaluation measures : We validate our method on the prostate MRI segmentation task. Where the data consist of six different sites that are collected from three public datasets with distribution shift, i.e. NCI-ISBI2013 [14], I2CVB [15], and Promise12 [16]. We evaluate the segmentation’s results through two evaluation metrics, Average Surface Distance (ASD) and the dice score (Dice).

Comparison with state-of-the-art generalization methods : We adopt the practical strategy, i.e. leave-one-domain-out. Thus, the model was trained on $k - 1$ distributed domains and tested on the one-left-out unseen domain. For the early stage of our work, we trained the model on each site separately by spilling the data as follows, i.e. 80%, 20% for the training and testing stages respectively. Then, we take up the most performant model that achieves the highest dice score, and use it as a pretrained model in the second stage. Where, we apply the DG strategy on the six sites as follows, for each experiment, we train the model using five sites as source domains and keep the left as one-left-out unseen target. We propose a comparative study of our work with recent DG-based prostate segmentation task, i.e. BigAug [7], which based on a stack of augmentation transformations, a meta-learning based method MASF [6], a shape-aware meta-learning SAML [11]. We present in table 1, the quantitative evaluation of the DG-methods aforementioned comparing to our results. We remark a significant improvement of our model on the sites 1, 3,4, and 6 comparing to the other methods. We further report the quantitative results in terms of dice score, where BigAug and MASF, SAML are more significantly performant comparing to the DeepALL. Our model performs well and improves over the baseline for Dice from 57.54% to 82.23% and over the BigAug, MASF, SAML in the site 4 by 93.01%. As well, we obtain a comparable results for the sites 1, 3, and 6. But, the dice score drops for the site 2 and 5.

Table 1. Generalization performance on prostate segmentation (Dice score (%) and ASD (pixel))

Method	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Average
DeepALL	60.36 7.24	47.06 5.11	78.10 1.45	41.96 8.64	47.59 3.50	70.22 2.47	57.54 4.74
BigAug [7]	88.62 1.70	86.22 1.56	83.76 2.72	87.35 1.98	85.53 1.90	85.83 1.75	86.21 1.93
MASF [6]	88.70 1.69	86.20 1.54	84.16 2.39	87.43 1.91	86.18 1.85	86.57 1.47	86.55 1.81
SAML [11]	89.66 1.38	87.53 1.46	84.83 2.07	88.67 1.56	87.37 1.77	88.34 1.22	87.67 1.58
Ours	82.96 0.31	79.54 1.21	85.96 0.66	93.01 0.20	70.74 3.82	81.22 2.47	82.23 1.44

4. Conclusion

To leverage the generalization capacity of the deep model, we introduced in this paper, the pretraining-finetuning learning scheme to mitigate the distribution shift and build a generalizable model that be able to perform on unseen data without prior knowledge for prostate MRI segmentation task. Experimental results show that the performance of our

work dropped in some cases and thus it affects the accuracy of prostate region detection. To deal with the drawbacks of our model, we seek in our future work, to enhance the learning scheme and investigate a novel loss function.

References

- [1] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annual review of biomedical engineering*. 2017 Jun 21;19:221-48.
- [2] Chen C, Dou Q, Chen H, Qin J, Heng PA. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*. 2020 Feb 10;39(7):2494-505.
- [3] Kamnitsas K, Baumgartner C, Ledig C, Newcombe V, Simpson J, Kane A, Menon D, Nori A, Criminisi A, Rueckert D, Glocker B. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: *International conference on information processing in medical imaging* 2017 Jun 25 (pp. 597-609). Springer, Cham.
- [4] Gibson E, Hu Y, Ghavami N, Ahmed HU, Moore C, Emberton M, Huisman HJ, Barratt DC. Inter-site variability in prostate segmentation accuracy using deep learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2018 Sep 16 (pp. 506-514). Springer, Cham.
- [5] Carlucci FM, D’Innocente A, Bucci S, Caputo B, Tommasi T. Domain generalization by solving jigsaw puzzles. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2019 (pp. 2229-2238).
- [6] Dou Q, Coelho de Castro D, Kamnitsas K, Glocker B. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*. 2019;32.
- [7] Zhang L, Wang X, Yang D, Sanford T, Harmon S, Turkbey B, Wood BJ, Roth H, Myronenko A, Xu D, Xu Z. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*. 2020 Feb 12;39(7):2531-40.
- [8] Zhou K, Yang Y, Hospedales T, Xiang T. Deep domain-adversarial image generation for domain generalisation. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 2020 Apr 3 (Vol. 34, No. 07, pp. 13025-13032).
- [9] Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. 2019 Jul 5.
- [10] Li H, Wang Y, Wan R, Wang S, Li TQ, Kot A. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*. 2020;33:3118-29.
- [11] Liu Q, Dou Q, Heng PA. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2020 Oct 4 (pp. 475-485). Springer, Cham.
- [12] Wang Z, Loog M, van Gemert J. Respecting domain relations: Hypothesis invariance for domain generalization. In: *2020 25th International Conference on Pattern Recognition (ICPR)* 2021 Jan 10 (pp. 9756-9763). IEEE.
- [13] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention* 2015 Oct 5 (pp. 234-241). Springer, Cham.
- [14] Bloch, N., Madabhushi, A., Huisman, H., Freymann, J., et al.: NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. (2015)
- [15] Lemaitre, G., Martí, R., Freixenet, J., Vilanova, J. C., et al.: Computer-Aided Detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. In: *Computers in Biology and Medicine*, vol. 60, pp. 8-31 (2015)
- [16] Litjens, G., Toth, R., Ven, W., Hoeks, C., et al.: Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. In: *Medical Image Analysis*, , vol. 18, pp. 359-373 (2014)

A Curated Dataset for Crack Image Analysis: Experimental Verification and Future Perspectives

Ammar M. OKRAN^{a,1}, Mohamed ABDEL-NASSER^{a,b}, Hatem A. RASHWAN^a and Domenec PUIG^a

^aDepartment of Computer Engineering and Mathematics, Rovira i Virgili University,
43007 Tarragona, Spain

^bElectrical Engineering Department, Aswan University, Egypt

Abstract. Most crack image datasets are developed for crack segmentation or detection. They cannot be used to train a deep learning model to detect and segment cracks simultaneously. Most of existing datasets do not include a very accurate annotation. Besides, some crack images cannot be used to train deep learning models because of their inferior quality. In this paper, we propose a promising curated crack image dataset that allows the development of crack segmentation, detection, and classification on the same set of images simultaneously. There is no dataset for road crack that involves detection and segmentation tasks to the best of our knowledge. The current version of the curated database consists of 506 images derived from the RDD2020 dataset taken from multi-countries (Japan, Czech, and India). We use the curated dataset to build different deep learning-based crack detection and segmentation methods. Our experiments demonstrate that the proposed dataset yields promising results for crack detection and segmentation.

Keywords. Road Crack, Deep learning, Mask-RCNN, Object detection, Instance Segmentation

1. Introduction

Human visual inspection of cracks is a time-consuming operation that also comes at a higher expense. As a result, computer vision-based crack detection systems can provide reliable results at a reasonable cost. Crack detection and segmentation remain a difficult challenge because of the following factors: lighting variations, poor continuity, low contrast between cracks and backdrop (e.g., pavement), intensity inhomogeneity on crack regions, and similar crack shadows. Developing efficient deep learning-based crack segmentation and detection models requires an extensive crack image dataset, including images captured under the aforementioned conditions. Table 1 summarizes the widely used datasets for crack detection and segmentation in the last 5 years. As one can see, these datasets are developed for crack segmentation or detection. They cannot be used to train a deep learning model to detect and segment cracks simultaneously. Also, most of these

¹Corresponding Author: E-mail: ammar.okran@urv.cat.

datasets do not include a very accurate annotation. Besides, some crack images cannot be used to train deep learning models because of their *very poor quality* (e.g., some images of RDD2020). There is *no dataset for road crack* that involves detection and segmentation tasks to the best of our knowledge.

In this paper, we propose a promising curated crack image dataset that allows the development of crack segmentation, detection, and classification on the same set of images simultaneously.

Table 1. Comparison between our curated crack image dataset and exiting ones.

Dataset	Year	Detection	Segmentation	Classification	Samples
GAPs v1 [1]	2017	✗	✗	✓	1969
RDD2018 [2]	2018	✓	✗	✓	9053
GAPs v2 [3]	2019	✓	✗	✓	2468
DeepCrack [4]	2019	✓	✗	✗	537
Angulo et al. [5]	2019	✓	✗	✓	18034
RDD2020 [6]	2020	✓	✗	✓	26620
Ours	2022	✓	✓	✓	506 (up-to-date)

2. Methodology

2.1. Data curation process

The curation process is shown in Figure 1. The curated dataset includes 506 images, carefully selected from the public dataset RDD2020 [6] which consists of 26,620 images of 3 countries (i.e., Japan, Czech, and India) taken by smartphone, with 850 road cracks and potholes. We selected images of acceptable quality and rejected the ones of bad quality. Also, when selecting the crack images, we tried keeping the same number of images per crack class (Longitudinal, Transverse, Alligator, and Pothole) to balance the dataset. An engineer has manually segmented each crack and drawn a bounding box around it. Quality control has been done on the generated annotations by AI engineers. Rejected annotations are repeated manually.

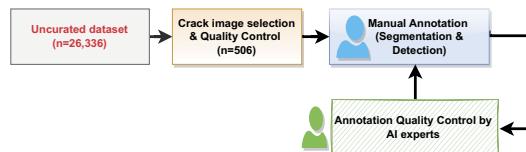


Figure 1. Data curation process.

2.2. Tested deep learning models for crack image analysis

In this study, we consider state-of-the-art deep learning models for crack segmentation and detection. In the task of *crack segmentation*, four segmentation models have been developed based on the Gated Skip Connections [7], High-Resolution Network (HR-Net), and Swin transformer [8], ConvNext [9], and the curated dataset. In turn, different *crack detection models* have been developed based on Mask R-CNN [10] and the curated dataset. Mask R-CNN [10] is a general framework for object instance segmentation that is simple to train, flexible, and adaptable. It recognizes objects in a photo while also creating a high-quality segmentation mask for each one.

3. EXPERIMENTS

Table 2 presents the results of different crack segmentation models trained and tested on the curated dataset. The developed crack segmentation models are HR-Net with fully connected layer (FC)—Fc+HRNet, UpperNet with Swin-Small as an encoder (UpperNet+Swin-Small), and UperNet with ConvNext-Base as an encoder (UpperNet+ConvNext-Base), and Gated skip connections with ConvNext-small as an encoder (Gated skip connections+ConvNext-small). As one can see, the Gated skip connections crack segmentation model achieves the best segmentation results. It obtains an IoU of 48.8 and a Dice score of 61.98. The results of the UperNet+ConvNext-Base model are close to the Gated skip connections model. UpperNet+Swin-Small achieves the best precision.

Table 2. Performance of the segmentation models on the curated dataset.

Model	Acc	IoU	Acc	Fscore	Precision	Recall	Dice
Fc+HRNet	98.71	42.74	47.58	55.27	69.7	47.58	55.27
UpperNet+Swin-Small	98.81	43.94	48.27	56.58	75.31	48.27	56.58
UpperNet+ConvNext-Base	98.71	48.48	55.1	61.84	72.61	55.1	61.84
Gated skip connections+ConvNext	98.8	48.8	55.59	61.98	72.08	55.59	61.98

Figure 2 shows quantitative results of the segmentation models, where each row shows a different crack—Longitudinal, Transverse, Alligator, and Pothole. As one can see, Fc+HRNet and UpperNet+Swin-Small under-segment the cracks. While both UpperNet+ConvNext and gated skip connections+ConvNext produce acceptable crack segmentation results in all shown examples. For crack detection, we develop different mod-

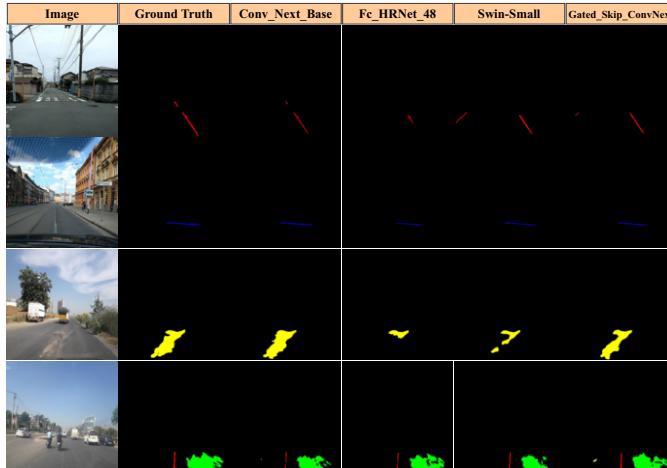


Figure 2. Segmentation results of different models. Each row presents a different crack—row1: Longitudinal Crack, row2: Transverse Crack, row3: Alligator Crack, and row4: Pothole.

els based on *Mask-RCNN*. Three models are shown here: standard *Mask-RCNN*, *Mask-RCNN_HRNet_32_1x*, and *Mask-RCNN_HRNet_40_1x*, where 32 and 40 stand for the width of the high-resolution convolution. It should be noted that the training of these models requires the bounding boxes of cracks and segmentation ground-truth masks, which are available in our curated dataset. Table 3 presents the results of the crack detec-

tion models trained on the curated dataset. Mask-RNN achieves a mean average precision (mAP) of 0.415 with the bounding box (i.e., *bbox_mAP*) and 0.392 with the segmentation branch (i.e., *segm_mAP*).

Table 3. Performance of the detection models on the curated dataset.

Model	bbox_mAP	segm_mAP
Mask-RCNN	0.415	0.392
Mask-RCNN.HRNet_32_1x	0.41	0.38
Mask-RCNN.HRNet_40_1x	0.402	0.394

4. Conclusion and Future work

This work introduced a promising curated crack image dataset that allows performing different crack image analyses simultaneously, like crack segmentation, detection, and classification, to be developed on the same collection of images. For crack segmentation and identification, we examined cutting-edge deep learning algorithms. Four segmentation models based on the Gated Skip Connections, High-Resolution Network, Swin transformer, ConvNext, and the curated dataset have been constructed for the task of crack segmentation. Future work will be focused on increasing the number of samples of the dataset and adding more crack classes.

References

- [1] Markus Eisenbach, Ronny Stricker, Daniel Seichter, Karl Amende, Klaus Debes, Maximilian Sesselmann, Dirk Ebersbach, Ulrike Stoekert, and Horst-Michael Gross. How to get pavement distress detection ready for deep learning? a systematic approach. In *2017 international joint conference on neural networks (IJCNN)*, pages 2039–2047. IEEE, 2017.
- [2] Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, and Hiroshi Omata. Road damage detection and classification using deep neural networks with smartphone images. *Computer-Aided Civil and Infrastructure Engineering*, 33(12):1127–1141, 2018.
- [3] Ronny Stricker, Markus Eisenbach, Maximilian Sesselmann, Klaus Debes, and Horst-Michael Gross. Improving visual road condition assessment by extensive experiments on the extended gaps dataset. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
- [4] Yahui Liu, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing*, 338:139–153, 2019.
- [5] Andres Angulo, Juan Antonio Vega-Fernández, Lina Maria Aguilar-Lobo, Shailendra Natraj, and Gilberto Ochoa-Ruiz. Road damage detection acquisition system based on deep neural networks for physical asset management. In *Mexican International Conference on Artificial Intelligence*, pages 3–14. Springer, 2019.
- [6] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Alexander Mraz, Takehiro Kashiyama, and Yoshihide Sekimoto. Deep learning-based road damage detection and classification for multiple countries. *Automation in Construction*, 132:103935, 2021.
- [7] M Jabreel and M Abdel-Nasser. Promising crack segmentation method based on gated skip connection. *Electronics Letters*, 56(10):493–495, 2020.
- [8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [9] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Reducing the Learning Domain by Using Image Processing to Diagnose COVID-19 from X-Ray Image

Maider ABAD^{a,1}, Jordi CASAS-ROMA^a and Ferran PRADOS^{a,b,c}

^a *e-Health Center, Universitat Oberta de Catalunya, Barcelona, Spain*

^b *Queen Square MS Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Science, University College of London, London, United Kingdom*

^c *Centre for Medical Image Computing (CMIC), Department of Medical Physics and Bioengineering, University College London, London, UK.*

Abstract. Over the last months, dozens of artificial intelligence (AI) solutions for COVID-19 diagnosis based on chest X-ray image analysis have been proposed. All of them with very impressive sensitivity and specificity results. However, its generalization and translation to the clinical practice are rather challenging due to the discrepancies between domain distributions when training and test data come from different sources. Consequently, applying a trained model on a new data set may have a problem with domain adaptation leading to performance degradation. This research aims to study the impact of image pre-processing on pre-trained deep learning models to reduce the learning domain. The dataset used in this research consists of 5,000 X-ray images obtained from different sources under two categories: negative and positive COVID-19 detection. We implemented transfer learning in 3 popular convolutional neural networks (CNNs), including VGG16, VGG19, and DenseNet169. We repeated the study following the same structure for original and pre-processed images. The pre-processing method is based on the Contrast Limited Adaptive Histogram Equalization (CLAHE) filter application and image registration. After evaluating the models, the CNNs that have been trained with pre-processed images obtained an accuracy score up to 1.2% better than the unprocessed ones. Furthermore, we can observe that in the 3 CNN models, the repeated misclassified images represent 40.9% (207/506) of the original image dataset with the erroneous result. In pre-processed ones, this percentage is 48.9% (249/509). In conclusion, image processing techniques can help to reduce the learning domain for deep learning applications.

Keywords. COVID-19, Transfer Learning, Deep Learning, Medical Image Processing, X-ray Imaging

¹Corresponding Author: Maider Abad, Universitat Oberta de Catalunya, Rambla del Poblenou, 156, 08018 Barcelona, Spain E-mail: mabdvz@uoc.edu.

1. Introduction

Coronavirus disease 2019 (COVID-19) is a novel illness caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). According to the World Health Organization, as of May 5th, 2022, there have been about 512M confirmed cases and 6M deaths [1]. Early in the pandemic, expectations were raised that chest CT or chest x-ray (CXR) might play a crucial role in the first-line diagnosis of COVID-19. Over time, the PCR test became more sensitive, and clinicians' understanding of the disease and how to treat it improved. However, the lab test still suffers from insufficient sensitivity, such as 71% reported in Fang et al. [2]. This is due to many factors, such as sample preparation and quality control [3]. Given the current sensitivity of the nucleic acid tests, many suspected patients must be tested multiple times several days apart before reaching a confident diagnosis. Hence, the imaging findings play a critical role in constraining the viral transmission and fighting against COVID-19 [4].

For this reason, over the last 24 months, there have appeared a lot of artificial intelligence (AI) solutions for COVID-19 CXR and CT diagnosis [5]. Compared to the traditional imaging workflow, which relies heavily on human labor, AI enables safer, more accurate, and efficient imaging solutions [4].

To improve diagnosis, Information Technology (IT) services from clinical health-care institutions have started to install a large number of the possible solutions in their networks, and they have integrated them with their image radiology viewers. In this way, radiologists can select one or more AI algorithms, send the data to compute, and get a diagnostic from each one. However, each of those solutions has been trained with different datasets. This difference in the training data increases the uncertainty of the correctness of any output of any AI solution. This issue is known as domain adaptation [6].

In addition, one of the main problems in evaluating clinical predictions is that deep learning models are often trained and tested on data that do not correspond to the target population and consequently makes it difficult for its generalization and application into a large scale setting. Therefore, a dataset may not adequately represent the range of possible patients and symptoms, generating a bias termed spectrum bias [7]. Because of this bias, most models that obtain high performance may perform poorly in other real-world scenarios where device vendor, acquisition parameters, image quality can vary.

Hence, the end-users of these tools, radiologists, feel overwhelmed with the impressive number of options that they can use from their radiology image viewers and the large number of different answers that they can get depending on the solution they choose.

However, several researchers are still working on developing reliable deep learning tools that can overcome the mentioned shortcomings. Li et al. [8] created a COVID-MobileXpert model for COVID-19 diagnosis on CXR images classified as COVID, Normal, and Pneumonia. They used 429 images for training (149 correspond to COVID-19) and 108 images for testing (36 of COVID-19), obtaining an accuracy of 88%.

Furthermore, Rahaman et al. [9] have compared different transfer learning approaches to identify COVID-19 samples from CXR images. They applied transfer learning to 15 models for 3-class classification (COVID-19, Normal, and Pneumonia). They used 720 images (220 COVID-19) for training and 140 (40 COVID-19) for testing. VGG19 (89.3% accuracy) and VGG16 (88.6% accuracy) showed the best results. Chowdhury et al. [10] trained and validated the ResNet101, MobileNetV2, CheXNet, SqueezeNet, and DenseNet201 models. They used 3,487 CXR images for 2-

class (COVID-19 and Normal) and 3-class classification (COVID-19, Normal, and Pneumonia). DenseNet201 models show the best performance achieving a sensitivity value of 99.7% and 97.9% for COVID-19 detection in the 2-class and 3-class classifications.

Other studies such as the one of Tahir et al. [11] and Rahman et al. [12] have used pre-processing techniques to evaluate the performance of different models. On the one hand, in [11], they have used the Contrast Limited Adaptive Histogram Equalization (CLAHE) filter, lung segmentation and image complement techniques to study its impact on the performance of the SqueezeNet, ResNet18, Inceptionv3, and DenseNet201 models for the classification of images into COVID-19, MERS, and SARS categories. The testing was done on 700 radiographs, of which 423 were COVID-19. The best results were obtained with Inceptionv3 reaching an accuracy of 98% for non-segmented CXR images. On the other hand, in [12], they have explored different image enhancement techniques for segmented and non-segmented CXR images. In this case, the best results were obtained with the ChexNet model and Gamma correlation method achieving an accuracy of 96% for non-segmented images.

In relation to this last study, Heidari et al. [13] have carried out an analysis of the performance of the VGG16 model by applying transfer learning for 3-class (Normal, Pneumonia, and COVID-19) recognition. They have used CXR images by applying diaphragm removal, histogram equalization, and bilateral low-pass filter pre-processing techniques. For the training, they used 8,474 radiographs (415 COVID-19) and 848 images (42 COVID-19) for testing. They obtained a weighted average precision of 95%, a recall of 94%, and an F1-score of 94%. All these studies have shown good performance for the diagnosis of COVID-19 and little generalization possibilities. We can observe that most of the datasets used for training and testing are unstructured, having difficulties over new learning domains for which they have not been trained.

This paper will analyze the impact of some image pre-processing techniques to reduce the learning domain from image-based datasets. For this purpose, we propose a comparison of the performance of unprocessed images and pre-processed images using three different models (VGG16, VGG19, and DenseNet-169) by applying transfer learning. The information in the manuscript is divided as follows: the gathered dataset used is introduced in Section 2.1. The processing techniques and the methodology are described in Sections 2.2 and 2.3, respectively. Section 2.4 presents the metrics used for the performance analysis. The results obtained with the proposed models and discussion are discussed in Section 3. Finally, in Section 4, the conclusions are summarized.

2. Methodology

2.1. Dataset

As a proof of concept, this research used 5,000 posterior anterior (PA) CXR images selected from 6 publicly available datasets to demonstrate the influence of image processing techniques in reducing the learning domain. The selection of the sub-sample taken from the different databases was done randomly. All the images were in PNG format. Of these 5,000 CXR radiographs, 4,000 belonging to the first 4 databases of Table 1 were used for training. 2,000 correspond to COVID-19 positive samples (COVID+), and the remaining 2,000 correspond to Normal or Non-Pathological healthy control (HC) samples. For the test, we used 1,000 images from the last two databases of Table 1.

Table 1. Summary of the datasets used in the research.

DB Name	Data Source	DB Sample Number	Sub-sample Number and Category	Reference
COVID-19 Data Repository	Institute for Diagnostic and Interventional Radiology, Hannover Medical School, Germany	243 COVID+	189 COVID+	[14]
COVID Radiography Dataset	Germany Medical School, SIRM, EURORAD, CXNet, COVID Chest XRay dataset, BIMCV and RSNA	3,616 COVID+ 1,345 Viral Pneumonia 6,012 COVID-Lung Opacity 10,192 HC	1,470 COVID+ 1,099 HC	[15]
COVIDGR	Hospital Universitario Clínico San Cecilio, Granada, Spain	426 COVID+ 426 HC	341 COVID+ 341 HC	[16]
Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification	University of California San Diego	4,273 Pneumonia 1,583 HC	560 HC	[17]
Covid ChestXray Dataset	Public sources, hospitals and physicians	506 Viral Pneumonia (468 COVID+) 46 Bacterial Pneumonia 26 Fungal Pneumonia 9 Lipoid Pneumonia 59 Unknown	468 COVID+ 532 HC	[18]
ChestX-ray8	Different hospitals' Picture Archiving and Communication Systems (PACS)	5,789 Atelectasis 1,010 Cardiomegaly 6,331 Effusion 10,317 Infiltration 6,046 Mass 1,971 Nodule 1,062 Pneumonia 2,793 Pneumothorax 84,312 HC		[19]

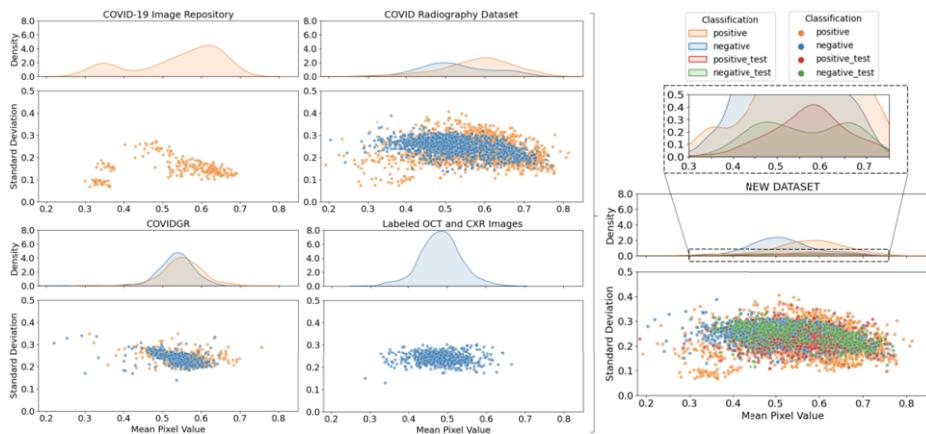


Figure 1. Analysis of the standard deviation and density distribution of the mean pixel value for each image of the datasets used (left) for training to generate the new dataset (right) including the training (in orange and blue) and test images (red and green).

Figure 1 shows the differences between the mean pixel values and the standard deviation of the images in each database and how the new database represents the unification of all the previous ones.

2.2. Data pre-processing

Subsequently, the next step involved pre-processing the incoming images using different techniques to compare the performance of the models with the original or non-preprocessed images and the pre-processed ones. The motivation for the application of pre-processing was, on the one hand, to improve the visual quality of the most damaged area of the lungs by increasing the image contrast through the Contrast Limited Adaptive Histogram Equalization (CLAHE) filter [20]. On the other hand, a rigid registration to an atlas of the lungs was performed for lung alignment and repositioning in the CXR radiographs to fix the deformed images. The atlas was created from the rigid registration

to a half-way space of 1,000 CXR images and the posterior averaging of all of them. Figure 2 shows the comparison between original and processed CXR images.

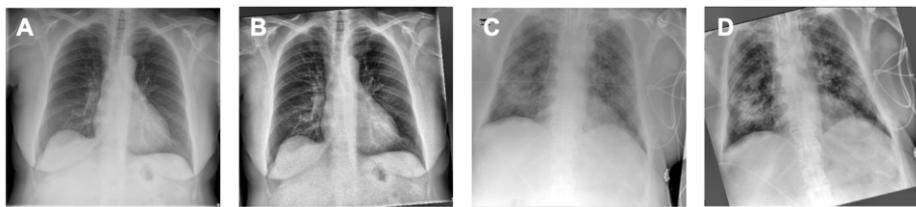


Figure 2. Sample of unprocessed COVID-19 positive case (A), unprocessed COVID-19 negative case (C) and their result after applying CLAHE filter and image registration (B,D).

2.3. Proposed method

Training models for disease detection requires a large data size, which sometimes may be difficult to achieve. To overcome the limited data size, we took advantage of pre-trained deep neural networks and transfer learning approaches. Pre-trained neural networks allow us to use a large number of parameters that have been previously trained on other datasets. Those parameters can be easily adapted to a new task with few modifications. Transfer learning is particularly beneficial in cases where there are not enough training samples to train a model from scratch, as may be the case for COVID-19 images.

We used transfer learning through ImageNet data which contains millions of labeled images. We applied this method to VGG16, VGG19 and DenseNet169 pre-trained deep neural networks to classify CXR images into negative and positive COVID-19 classes.

In addition, all images in the dataset were normalized between 0 and 1 and resized to 480 x 480 pixels. In the case of using the unprocessed database, no further modifications were applied to the images. In the case of the preprocessed images, we realigned all the images to an in-house lung atlas template using a rigid transformation and afterwards applied a CLAHE filter [20] to automatically enhance the contrast of each image.

We took the architectures of the trained CNNs with ImageNet, which includes 1,000 class lists, for transfer learning [21]. The set of features learned by these networks was used by transfer learning to extract specific features related to COVID-19 detection. The weights of each layer of the CNN models were frozen, except for those of the fully connected output layer. This layer was removed to adapt the model to the new classification problem. The default hyperparameters for each of the networks were maintained.

We added six layers on the top of each model (Figure 3). First, we set a 4x4 Average Pooling layer to take the average value of the features from the feature maps, and thus reduce computation complexity and variance avoiding overfitting. Next, we added a Flatten layer to convert the multi-dimensional data into 1-dimensional array. Then, we attached two fully connected layers with an output of 128 and 64, respectively, using Rectified Linear Unit (ReLU) activation function. ReLU, which is defined as: $\text{ReLU}(x) = \max(0,x)$, is applied after each of those layer to make the network non-linear. We also add a dropout layer with a rate of 0.5 to avoid overfitting. Finally, another fully connected layer with a softmax classifier and an output of two classes corresponding to negative and positive COVID-19 detection. Softmax assigns each node a number between 0 and

1, being the sum of all probabilities equal to 1. To train the CNN models, we set the batch size to 10 and the number of epochs to 20. We used the Adam optimizer with a learning rate of 0.0001 for optimization. Adam optimization is a stochastic gradient descent method that involves a combination of first and second-order moments. In this work, we employed categorical cross-entropy as the loss function. Moreover, we used early stopping as a kind of regularization used to avoid overfitting when training. The criterion we have used is based on stopping training when the validation loss increases noticeably.

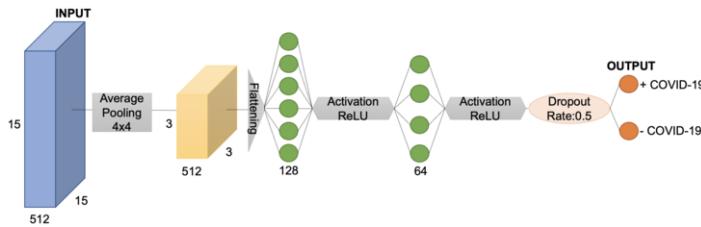


Figure 3. Transfer learning model scheme.

The training dataset (databases 1-4 of Table 1) was split using 10% of the images for validation and 90% for training. In addition, the partition was 50% COVID+ and 50% HC and all datasets were included in the training and validation groups. Test has been done with an independent dataset (databases 5-6 of Table 1) that has not been used for training. The partitioning of the testing dataset is 46.8% COVID+ and 53.2% HC.

2.4. Performance analysis

We evaluated true positives (TP) and true negatives (TN) to represent images that were well classified as COVID+ or HC respectively, while false positives (FP) and false negatives (FN) represent the misclassified images. The performance was measured in terms of sensitivity, specificity, accuracy, F1-score, and precision.

3. Results and Discussion

By testing the unprocessed and pre-processed images and applying the metrics mentioned in the performance analysis, we obtained the results shown in Table 2. We observe an improvement in accuracy (+1.1%) for the pre-processed images (82.5%) compared to the unprocessed ones (81.4%) for VGG16 model. In addition, we see that the sensitivity improves notoriously (+4.8%) for VGG16 model. This value is crucial for COVID-19 detection as it indicates a decrease in FN values (see Figure 4). As for the specificity, we observe that the value decreases (by -2.6%) in the pre-processed images because of the slight increase in the FN values, as shown in the confusion matrices (Figure 4). Moreover, for this reason, the precision value is reduced by 1.0% in the pre-processed images, and the F1-score improves by 1.7%.

The VGG19 model testing results do not show the same pattern as in the previous model presented in Table 2 and Figure 4. In this case, the accuracy of the pre-processed images decreases up to 82.6% (by -0.6%) as an F1-score 82.2% (by -1.0%), and sensitiv-

Table 2. Comparison between unprocessed and processed images for the three networks in terms of accuracy, precision, sensitivity, specificity and F1-score (%). The best results are shown in bold. All the metrics results are reported considering the 95% confidence interval.

CNN Model	Dataset	Accuracy	Specificity	Sensitivity	F1-score	Precision
VGG16	Original	81.4 ± 2.4 CI[79.0 to 83.8]	79.4 ± 3.4 CI[76.0 to 82.8]	83.4 ± 3.4 CI[80.0 to 86.8]	81.7 ± 3.5 CI[78.2 to 85.2]	80.2 ± 3.6 CI[76.6 to 83.8]
	Pre-processed	82.5 ± 2.4 CI[80.1 to 84.9]	76.8 ± 3.6 CI[73.2 to 80.4]	88.2 ± 2.9 CI[85.3 to 91.1]	83.4 ± 3.4 CI[80.0 to 86.8]	79.2 ± 3.7 CI[75.5 to 82.9]
VGG19	Original	83.2 ± 2.3 CI[80.9 to 85.5]	83.0 ± 3.2 CI[79.8 to 86.2]	83.4 ± 3.4 CI[80.0 to 86.8]	83.2 ± 3.4 CI[79.8 to 86.6]	83.1 ± 3.4 CI[79.7 to 86.5]
	Pre-processed	82.6 ± 2.3 CI[80.3 to 84.9]	85.0 ± 3.0 CI[82.0 to 88.0]	80.2 ± 3.6 CI[76.6 to 83.8]	82.2 ± 3.5 CI[78.7 to 85.7]	84.2 ± 3.3 CI[80.9 to 87.5]
DenseNet-169	Original	82.8 ± 2.3 CI[80.5 to 85.1]	82.0 ± 3.3 CI[78.7 to 85.3]	83.6 ± 3.4 CI[80.2 to 87.0]	82.9 ± 3.4 CI[79.5 to 86.3]	82.3 ± 3.5 CI[78.8 to 85.8]
	Pre-processed	84.0 ± 2.3 CI[81.7 to 86.3]	77.6 ± 3.5 CI[74.1 to 81.1]	90.4 ± 2.7 CI[87.7 to 93.1]	85.0 ± 3.2 CI[81.8 to 88.2]	80.1 ± 3.6 CI[76.5 to 83.7]

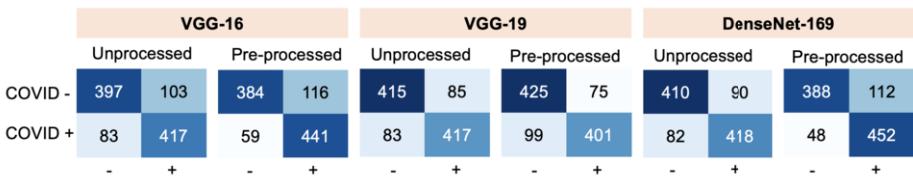


Figure 4. The figure shows the confusion matrix results after testing the three models. Positive cases are indicated as COVID+ or “+”, and negative cases as COVID- or “-”. Rows represent the true label and columns represent the predicted label.

ity to 66.2% (by -3.2%). However, specificity increases by 2.0% and precision by 1.1%. This is related to the results of the confusion matrices in Figure 4.

Regarding the test results for DenseNet-169 model (Table 2), we see an increase in accuracy from 82.8% to 84.0% for the pre-processed images. Similarly, the F1-score increases to 85.0% for the pre-processed images and the sensitivity to 90.4%. In this case, specificity decreases by 4.8% and precision by 2.2%. In this case, the confusion matrices of Figure 4 show a notorious reduction of 41.5% in the FN values of the pre-processed images. Consequently, there is a 24.4% increase in the FP values.

Overall, the best results have been obtained for the pre-processed images, which is related to the accuracy and sensitivity increase. This last point is remarkable since it indicates that the number of images classified as FN, which represent COVID+ cases classified as HC, is reduced.

In addition, we performed a study to compare the images classified as FN and FP considering the three models (VGG16, VGG19, and DenseNet-169) for the two datasets (original and preprocessed). We found that the repeated misclassified images represent 40.9% of the original image dataset with the erroneous result since 207 repeated images (69 images for each CNN model) were present in a total of 506 misclassified unprocessed images for the three models. This percentage is 48.9% in pre-processed ones since 249 repeated images (83 images for each CNN model) were present in a total of 509 misclassified pre-processed images.

Considering that a large part of the misclassified images are repeated, we wanted to check if there is a pattern capable of differentiating these images from those that have been well classified (TN and TP) by the three models in each dataset. To do this, we first studied the mean pixel values and the standard deviation of each group. Figure 5 shows that the mean pixel values of the TN images are clustered between 0.40 and 0.75 on a scale of 0 to 1. While the TP images extend equally over a wide range in the case of the unprocessed images, the pre-processed images show a density peak between 0.50 and 0.65 values. FN and FP values are under the spectrum of the well classified, so it is hard to draw a clear differentiation.

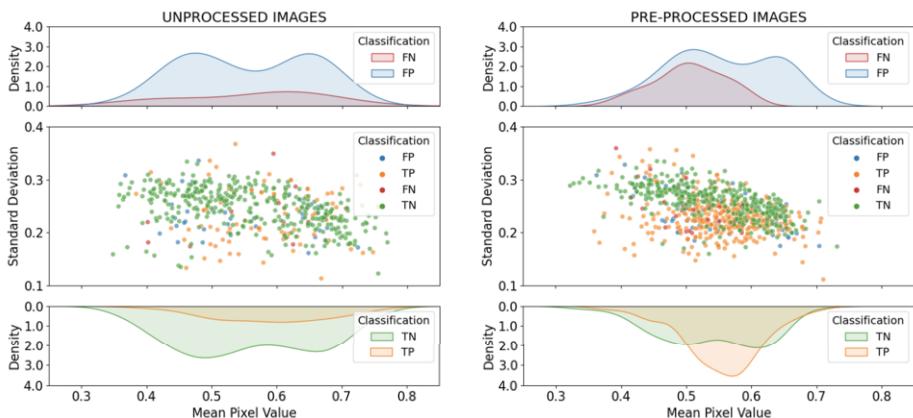


Figure 5. Representation of the standard deviation and density of the mean pixel values for each image classified as TP, TN, FP and FN.

With the purpose of looking for differentiation between misclassified and well-classified images, we carried out an entropy's study to compare both groups. Within this study, we have considered as FN and FP only images equally misclassified by the three models. The groups TP and TN correspond to the images equally well classified as TP or TN by the three models. Shannon Entropy [22] is a measure from information theory which represents the level of uncertainty of a set of values and we employed it to characterize the overall image's texture coherence.

Figure 6 shows the results of this study, where we can observe a clear differentiation between the misclassified images, which have higher entropy values, and the well-classified ones. This difference could serve as an initial filter for identifying these images before using them in convolutional networks.

Furthermore, we studied contrast, signal-to-noise ratio (SNR) and image normalized entropy parameters to detect similarities between misclassified and well-classified images. To this end, after calculating those parameters for each image, we applied the t-test to check whether the groups were statistically different or not. The significance level has been set to $p < 0.05$. Table 3 shows the results of the different p-values obtained in the tests considering as FN and FP the images misclassified in common by the three models at each dataset, and TP and TN as the well classified images in common by the three models at each dataset. The results indicate that FN images are misclassified because their resemblance to the TN images in terms of SNR and contrast.

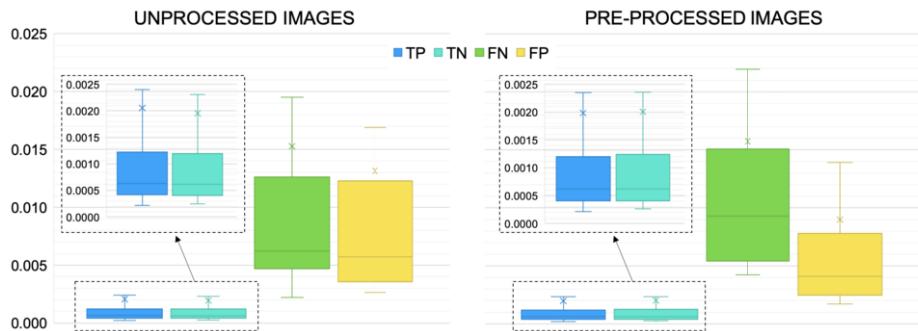


Figure 6. Representation of the Shannon entropy for each image classified as TP, TN, FP and FN.

Table 3. Results of the p-values obtained from the t-test.

Metric	Dataset	FN-TN	FN-TP	FP-TN	FP-TP
SNR	Original	0.865	<0.001	0.215	0.004
	Pre-processed	0.464	<0.001	0.002	0.447
Contrast	Original	0.954	<0.001	<0.001	0.017
	Pre-processed	0.286	<0.001	0.690	<0.001
Entropy	Original	0.006	0.006	0.005	0.005
	Pre-processed	0.004	0.004	0.003	0.002

4. Conclusion

In the present work, we conducted a comparative study of three CNN models trained and tested with unprocessed and pre-processed CXR images to detect COVID-19. For this purpose, we have created a dataset using 5,000 images from various publicly available sources. After evaluating different classic metrics such as accuracy, sensitivity, or specificity, we have found that in most cases accuracy and sensitivity improve when using pre-processed images. We found that these pre-processing steps are directly related to the decrease of images wrongly classified as non-COVID-19. The CNN model DenseNet-169, trained and tested on pre-processed images, achieved an accuracy of 84.0%, a sensitivity of 90.4%, and a specificity of 77.6%, showing the best performance results in terms of positive COVID-19 detection. In addition, we have demonstrated that many misclassified images show higher entropy values than the well-classified ones, so our future work will focus on introducing an entropy filter for improving the performance of the prediction models. Moreover, we would like to implement other processing techniques such as those in [11,12,13] and thus test the change in learning domain reduction in both the models used in this article and in others mentioned in [11,12,9]. In conclusion, this study has shown that well selected image processing techniques can help to reduce the learning domain for deep learning applications.

References

- [1] WHO Coronavirus (COVID-19) Dashboard;. Accessed: 2022-5-9. <https://covid19.who.int>.

- [2] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al.. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR; 2020.
- [3] Rahmani AM, Mirmahaleh SYH. Coronavirus disease (COVID-19) prevention and treatment methods and effective parameters: A systematic literature review. Sustain Cities Soc. 2021 Jan;64:102568.
- [4] Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Rev Biomed Eng. 2021 Jan;14:4-15.
- [5] Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M. Mapping the landscape of Artificial Intelligence applications against COVID-19; 2020.
- [6] Choudhary A, Tong L, Zhu Y, Wang MD. Advancing Medical Imaging Informatics by Deep Learning-Based Domain Adaptation. Yearb Med Inform. 2020 Aug;29(1):129-38.
- [7] Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022 Apr;5(1):48.
- [8] Li X, Li C, Zhu D. COVID-MobileXpert: On-Device COVID-19 Patient Triage and Follow-up using Chest X-rays; 2020.
- [9] Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. J Xray Sci Technol. 2020;28(5):821-39.
- [10] Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, et al.. Can AI Help in Screening Viral and COVID-19 Pneumonia?; 2020.
- [11] Tahir AM, Qiblawey Y, Khandakar A, Rahman T, Khurshid U, Musharavati F, et al. Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images. Cognit Comput. 2022 Jan:1-21.
- [12] Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Kashem SBA, et al.. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images; 2021.
- [13] Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms; 2020.
- [14] ml-workgroup. GitHub - ml-workgroup/covid-19-image-repository: Anonymized dataset of COVID-19 cases with a focus on radiological imaging. This includes images (x-ray / ct) with extensive metadata, such as admission-, ICU-, laboratory-, and patient master-data;. Accessed: 2022-5-9. <https://github.com/ml-workgroup/covid-19-image-repository>.
- [15] Rahman T. COVID-19 Radiography Database;.
- [16] Tabik S, Gomez-Rios A, Martin-Rodriguez JL, Sevillano-Garcia I, Rey-Area M, Charte D, et al. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. IEEE J Biomed Health Inform. 2020 Dec;24(12):3595-605.
- [17] Kermany D. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley Data; 2018.
- [18] ieee8023. IEEE8023/covid-chestxray-dataset: We are building an open database of COVID-19 cases with chest X-ray or CT images.; Available from: <https://github.com/ieee8023/covid-chestxray-dataset>.
- [19] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 July.
- [20] Asnaoui KE, El Asnaoui K, Chawki Y, Idri A. Automated Methods for Detection and Classification Pneumonia Based on X-Ray Images Using Deep Learning; 2021.
- [21] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database; 2009.
- [22] Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(3):379-423.

On the Importance of Color Pre-Processing for Object Detection in Submarine Images

Jose-Luis LISANI ^{a,1}, Ana Belén PETRO ^a, Catalina SBERT ^a,
Amaya ÁLVAREZ-ELLACURÍA ^b, Ignacio A. CATALÁN ^b, and Miquel PALMER ^b

^aUniversitat Illes Balears (UIB), Spain

^bIMEDEA (UIB-CSIC), Spain

Abstract. When training a neural network for object detection a great deal of effort is usually devoted to augment the training dataset. The rationale behind this process is that augmentation increases the generalization capability of the network. However, little attention has been paid to the application of image enhancement techniques as a pre-processing step of the training task. In this paper we show, in the context of fish detection in submarine images, that the application of classical color enhancement methods may improve significantly the performance of the well known Mask R-CNN object detector.

Keywords. image enhancement, object detection, CNN, submarine images

1. Introduction

The use of image enhancement techniques is common in many vision tasks. These techniques are used to improve the contrast, the brightness and the color of the images, usually as a pre-processing step to a further analysis. As an example, we see in Figure 1 one original subaquatic image and the result of its processing with the Retinex algorithm [1]. We observe that the objects in the scene (in particular the fish) are more easily distinguishable in the processed image than in the original. In general, enhancement helps humans in detection tasks. The question arises whether the same is true for deep learning-based detection algorithms. In this paper we seek to find an answer to this question in the particular case of fish detection in submarine images.

Some recent papers [2,3] have shown that the visual quality of an image is not necessarily correlated with the accuracy of an object detector that uses this image as input. We study in the current work how the performance of a popular CNN for object detection (Mask R-CNN [4]) is affected by the use of five representative underwater image enhancement algorithms.

The paper is organized as follows. Section 2 gives a short overview of the techniques used for the enhancement of underwater images and presents the five algorithms selected for the study. Section 3.1 describes the CNN used in the experiments, the set of images used for training and evaluation, and details the training parameters. In Section 3.2 an statistical analysis of the results is provided. Finally, some conclusions are drawn in Section 4.

¹Corresponding Author: Jose-Luis Lisani, Universitat Illes Balears, Spain; E-mail: joseluis.lisani@uib.es



Figure 1. Original image (left) and result of processing with Retinex algorithm [1] (right).

2. Enhancement of Underwater Images

Underwater images suffer from color cast, low contrast and haze due to the different attenuation of the light wavelengths and to the scattering effect. These effects are depth-dependent.

Model-free or prior-based methods can be used to enhance these images. The former seek to improve the visual quality without taking the depth dependency into account while the later are based on physical image-formation models. Model-free methods are simple and applicable to a wider type of images, but are not always able to correctly improve them. Prior-based approaches do not always obtain good results due to the use of over-simplified models or to the difficulty to estimate correctly the parameters of the model in a general case. In recent years, a new trend of enhancement methods has emerged, the deep-based (or data-driven) approaches. However, the lack of available training data limits its performance [5].

In our study we have selected two popular model-free methods (MSR [1] and Fusion [6]), and three prior-based methods (UDCP [7], ARC [8] and InfoLoss [9]). MSR (Multi-Scale Retinex) aims at removing global illumination changes by locally improving the contrast of the image. Fusion combines contrast enhanced and color corrected versions of the original image using a multi-scale strategy. Both UDCP (Underwater Dark Channel Prior) and ARC (Automatic Red-Channel) estimate the depth map of the image and use this information to restore the color balance, but while UDCP bases its estimation on the green and blue color channels, ARC uses the red channel. Finally, InfoLoss consists of two steps, first a dehazing method is applied after estimating the depth map of the image, and then a contrast enhancement algorithm is applied.

We have used our own implementations of MSR [10], Fusion, UDCP and InfoLoss, based on the descriptions provided in the original papers. For ARC we have used the online tool for underwater image processing <https://puiqe.eecs.qmul.ac.uk/>.

3. Experiments

We have trained a popular CNN for object detection (Mask R-CNN [4]) using different processed versions of the same original images. We have then compared the mAP values obtained with the trained networks on a common test set.

Mask R-CNN permits simultaneous detection, classification and instance segmentation of the image objects. The network consists of a Backbone for feature extraction (we use Resnet101 in our tests), a Region Proposal Network (RPN) and three output branches, for bounding box location, object classification and segmentation,

respectively. We have used the implementation of the network available at https://github.com/matterport/Mask_RCNN.

3.1. Experimental Setting

We have collected a dataset of 600 underwater images, coming from two different locations in the Mallorcan coast. All the fish in these images have been manually segmented and the network has been trained to detect them. The dataset has been split into three sets: a training set (400 images, with 4252 annotated fish), a validation set (100 images, 917 fish, used for tuning the hyperparameters of the training) and a test set (100 images, 1004 fish, for evaluation of the results).

Six versions of the dataset have been used in the experiment: the original images and also the images processed with the five methods described in the previous section. The CNN has been trained and evaluated using these six datasets.

The following training strategy has been used: the upper layers of the network ('heads') have been trained for 30 epochs; the intermediate layers (fc3) have been trained for 30 additional epochs; finally, all the layers have been trained for other 30 epochs. In order to reduce the effect of the random nature of the minimization process in the obtained results, the above strategy has been repeated five times, and the mean average precision values (mAP) obtained on the test sets have been recorded.

3.2. Statistical Analysis of the Results

Figure 2 displays the mAP values obtained on the test set for each one of the trained networks, using as input the images pre-processed with the different methods.

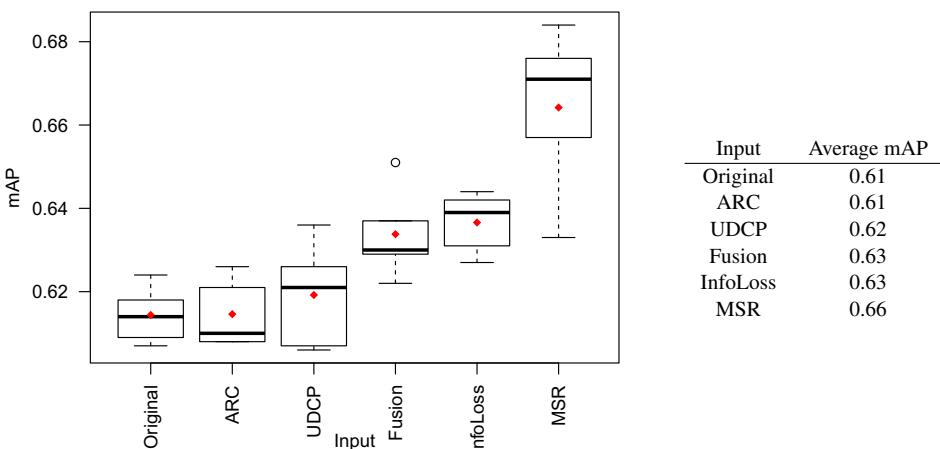


Figure 2. Boxplots of mAP values on the test set, obtained after pre-processing the input images with different enhancement techniques. The table displays average mAP values, represented as red dots in the figure.

Visually one can observe that almost all pre-processing methods improve the performance obtained with the un-processed images. In particular, MSR obtains a 5% increase on average.

In order to check if these differences are statistically relevant we perform a one way ANOVA test (Analysis of Variance) to the obtained values. The ANOVA result is signif-

icant ($F = 12.79, df = 5, p < 0.0001$), thus at least one group is significantly different from the rest. Additionally, we apply the Tukey Honest Significant Differences post-hoc test to obtain pairwise comparisons of the methods. The results (p-values) of the test are displayed in Table 1. We observe that significant differences are obtained when comparing MSR with the rest.

	Original	ARC	UDCP	Fusion	InfoLoss	MSR
Original	-	1.0	0.99	0.14	0.07	< 0.001
ARC	-	-	0.99	0.15	0.07	< 0.001
UDCP	-	-	-	0.41	0.23	< 0.001
Fusion	-	-	-	-	0.99	< 0.01
InfoLoss	-	-	-	-	-	< 0.05

Table 1. P-values corresponding to the pairwise comparison of the methods using the Tukey test. Statistically meaningful differences are marked in bold type.

4. Conclusions

The obtained results show that the performance of an object detection network can be increased by preprocessing the input images (both during the training and the inference steps) using classical enhancement methods. In particular, for the case of underwater images, the use of the Multi-Scale Retinex method permits to significantly increase the mAP value by a 5% on average, with respect to the original un-processed images. As a continuation of this work we shall investigate how the use of augmentation techniques, both on the original and pre-processed images, may affect the detection results.

Acknowledgements

This work has been sponsored by the Comunitat Autònoma de les Illes Balears through the Direcció General de Política Universitària i Recerca with funds from the Tourist Stay Tax Law ITS 2017-006 (PRD2018/26).

References

- [1] Jobson DJ, Rahman Z, Woodell GA. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Trans on Image Processing*. 1997.
- [2] Liu R, Fan X, Zhu M, Hou M, Luo Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans on Circuits and Sys for Video Tech*. 2020;30(12):4861-75.
- [3] Chen L, Jiang Z, Tong L, Liu Z, Zhao A, Zhang Q, et al. Perceptual Underwater Image Enhancement With Deep Learning and Physical Priors. *IEEE Trans on Circuits and Sys for Video Tech*. 2021;31(8):3078-92.
- [4] He K, Gkioxari G, Dollr P, Girshick R. Mask R-CNN. In: *IEEE Int. Conf. on Computer Vision*; 2017. p. 2980-8.
- [5] Anwar S, Li C. Diving deeper into underwater image enhancement: A survey. *Signal Processing: Image Communication*. 2020;89:115978.
- [6] Ancuti CO, Ancuti C, C DV, Bekaert P. Color balance and fusion for underwater image enhancement. *IEEE Trans on Image Processing*. 2018;27(1):379-93.
- [7] Drews Jr P, do Nascimento E, Moraes F, Botelho S, Campos M. Transmission Estimation in Underwater Single Images. In: *2013 IEEE Int. Conf. on Computer Vision Workshops*; 2013. p. 825-30.
- [8] Galdran A, Pardo D, Picón A, Alvarez-Gila A. Automatic Red-Channel underwater image restoration. *Journal of Visual Communication and Image Representation*. 2015;26:132-45.
- [9] Li CY, Guo JC, Cong RM, Pang YW, Wang B. Underwater Image Enhancement by Dehazing With Minimum Information Loss and Histogram Distribution Prior. *IEEE Trans on Image Processing*. 2016;25(12):5664-77.
- [10] Petro AB, Sbert C, Morel JM. Multiscale Retinex. *Image Processing On Line*. 2014;71-88.

Referenceless Image Quality Assessment Utilizing Deep Transfer-Learned Features

Basma AHMED¹, Osama A. OMER³, Amal RASHED¹ and Domenec PUIG²,
Mohamed ABDEL-NASSER^{2,3}

¹*Faculty of Computers and Information, South Valley University, Qena, Egypt*

²*Computer Engineering and Mathematics Department, University Rovira i Virgili,
Tarragona, Spain*

³*Electrical Engineering Department, Aswan University, Aswan, Egypt*

Abstract. Image quality assessment (IQA) algorithms are critical for determining the quality of high-resolution photographs. This work proposes a hybrid NR IQA approach that uses deep transfer learning to enhance classic NR IQA with deep learning characteristics. Firstly, we simulate a pseudo reference image (PRI) from the input image. Then, we used a pre-trained inception-v3 deep feature extractor to generate the feature maps from the input distorted image and PRI. The distance between the feature maps of the input distorted image and PRI are measured using the local structural similarity (LSS) method. A nonlinear mapping function is used to calculate the final quality scores. When compared to previous work, the proposed method has a promising performance.

Keywords. Blind image quality, Similarity measures, Pseudo-reference, Deep learning

1. Introduction

Several quality degradations are introduced in each phase of the visual communication system, e.g., capturing, transmission, compression, and display. However, high-fidelity images are needed in many fields, e.g., remote sensing image recognition, virtual reality, medical imaging, and other fields. Image quality assessment (IQA) algorithms can quantify the quality of visual content delivered to end-users, which can be adopted as the quantify criteria or optimization goal embedded in the visual communication systems [1]. IQA methods can be split into subjective and objective, according to the need for human eyes for ranking [2]. Objective assessment is more feasible and extensively applied because a machine can automatically forecast the quality of an image utilizing mathematical models. Objective assessment can be generally divided into three categories according to the presence or deficiency of a reference image: 1) full-reference IQA (FR-IQA), 2) reduced-reference IQA (RR-IQA), and 3) no-reference or blind IQA (BIQA) [14]. A reference image is often not given in practical use processes, which hinders the application remit of FR-IQA and RR-IQA. This study is focused on NR IQA. In the literature, many NR IQA has been proposed. For instance, Le et al. [6] suggested a blind technique to foretell the visual quality of multiply distorted images based on structural degradation. A structural feature is extracted as the gradient-weighted histogram of the local binary pattern (LBP) studied

on the gradient map. They also utilized LBP extracted from texture and structural maps. Q. Wu et al. [7] suggested a local pattern statistics index (LPSI) mechanism by selecting statistical features extracted from binary patterns of local image structures. Min et al. [8] proposed a blind image quality assessment based on a pseudo reference image, and they used a “reference” called pseudo reference image (PRI) to assess blockiness, sharpness, and noisiness. Jinbin Hu et al. [1] proposed a deep network-based blind image quality assessment utilizing two-side pseudo reference images. In turn, deep convolutional neural networks (CNNs) are the most successful architectures for image analysis. Deep transfer learning makes it possible to use previously learned features from one learning task to another learning task as initial features. Kang et al. [9] suggested a CNN to assess image quality without a reference image.

By defining a *pseudo image* as a benchmark, the problem of NR IQA can be achieved via a pseudo reference (PR) IQA solution to bridge a practical approach between FR and NR IQAs. In such a solution, a perfect undistorted picture is not required and may not possibly be available. Besides, we can conclude that transfer learning can extract robust quality-aware features from the image without needing labelled data. This paper proposes a hybrid NR IQA method that enriches traditional NR IQA with deep learning features via deep transfer learning. The method first generates a PRI from the input distorted image. A pre-trained CNN feature extractor has been employed in our study to generate feature maps from the input distorted image and its PRI. We use the local structural similarity (LSS) to measure the distance between the feature maps extracted from the input distorted image and a PRI generated from the input distorted image. A nonlinear mapping function is used to compute the final quality scores.

2. Methodology

Figure 1 presents the proposed NR IQA method. The main components of the proposed method are the simulation of PRI, extracting deep learning features, generation of feature maps, calculation of the similarity between feature maps of the input distorted image and PRI, and applying nonlinear mapping on the similarity scores to compute the final NR IQA score for the input distorted image.

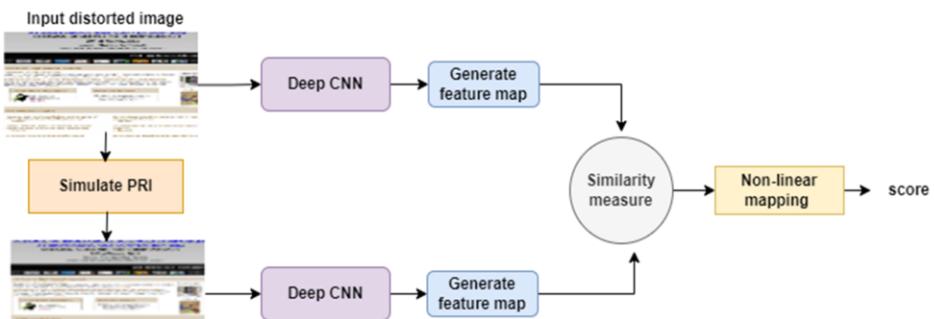


Figure 1. Proposed NR IQA method.

Following the fact that excessive blurring introduces pseudo structures that can be used to judge the quality of the blurred image. In this study, we use a 3×3 averaging filter to derive the PRI. Given a distorted image I , PRI is computed as follows:

$$\text{PRI} = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} * I \quad (1)$$

We utilize the pre-trained inception-v3 deep CNN that is 48 layers deep, trained on more than a million images from the ImageNet database to extract features from images. We use the local structural similarity (LSS) to measure the similarity between the feature map of the distorted image I and the corresponding PRI feature map. The local structure maps of the distorted image I and the PRI can be described as $K_d = (k_{d^{ij}})_{h \times w}$, $K_m = (k_{m^{ij}})_{h \times w}$ respectively, where h, w denote the rows and columns of the image.

We determine the overlap between K_d, K_m as follows:

$$k_{0^{ij}} = (k_{d^{ij}} \cdot k_{m^{ij}})_{h \times w}, \quad (2)$$

We can also define the union between them as follows:

$$k_{u^{ij}} = (k_{d^{ij}} \setminus k_{m^{ij}})_{h \times w} \quad (3)$$

Then, LSS can be defined as follows:

$$LSS = \frac{\sum_{i,j} k_{0^{ij}}}{\sum_{i,j} k_{u^{ij}} + 1} \quad (4)$$

We employ a five-parameter logistic function as a non-linear mapping function to map the quality scores, which can be defined as follows:

$$q' = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2(q - \beta_3))} \right) + \beta_4 q + \beta_5, \quad (5)$$

where q' and q stand for the original and mapped quality scores, respectively; $\{\beta_j | j = 1, 2, \dots, 5\}$ are five parameters determined through curve fitting. In the literature, the q' values are considered for evaluation metrics computation.

Two IQA databases are adopted as testing platforms, SIQAD [13] and CSIQ [12]. All datasets consist of numerous subsets of different distortion types. In this work, we focus on Gaussian blurring. The CSIQ dataset contains 30 original screen images and 150 Gaussian blur images. The SIQAD dataset contains 20 original screen images and 140 Gaussian blur images. In this study, three performance indexes are adopted to measure the proposed method:

(1) Pearson Linear Correlation Coefficient (PLCC):

$$\text{PLCC} = \frac{1}{n-1} \sum_{j=1}^n \left(\frac{x_j - \bar{x}}{\sigma_x} \right) \left(\frac{y_j - \bar{y}}{\sigma_y} \right) \quad (6)$$

where $\{x_1, x_2, \dots, x_n\}, \{y_1, y_2, \dots, y_n\}$ stand for a MOS value and a predicted one, respectively, \bar{x} and \bar{y} are their average scores, and σ_x and σ_y are their variances.

(2) Spearman's Rank Ordered Correlation Coefficient (SROCC):

$$\text{SROCC} = 1 - \frac{6}{n(n^2 - 1)} \sum_{j=1}^n (r_{x_j} - r_{y_j})^2, \quad (7)$$

Where r_{x_j} and r_{y_j} represent the rank of x_j and y_j in MOS values and predicted ones, respectively.

(3) Root-Mean-Square Error (RMSE):

$$\text{RMSE} = \left[\frac{1}{n} \sum_{j=1}^n (x_j - y_j)^2 \right]^{\frac{1}{2}}, \quad (8)$$

RMSE is a metric that measures the absolute error between the subjective and objective scores.

3. Preliminary results

We compared the performance of the proposed model with NR IQA models such as DIIVINE [4], BLIINDS-II [10], BRISQUE [3], CORNIA [11], NIQE [5]. As shown in Table 1, the proposed method achieves PLCC, SROCC and RMSE values of 0.7117, 0.6035, and 8.0729 with the SIQAD dataset, respectively. From this table, we can see that the proposed model is comparable to the best-performed metrics. With the CSIQ dataset, Table 2 shows that BRISQUE achieves PLCC and SROCC values of 0.9275 and 0.9026. The proposed method achieves 0.1015 RMSE, better than DIIVINE, BLIINDS-II, BRISQUE, CORNIA, NIQE metrics. To test the proposed method in predicting the quality of images with Gaussian blur GB, we compare the predicted score and ground-truth DMOS scores in Fig.2. The deviation between the DMOS and predicted scores is relatively small, reflecting high predictors in agreement with human visual perception.

Table 1. Comparison with existing IQA algorithms for SIQAD dataset.

Algorithm	PLCC	SROCC	RMSE
DIIVINE	0.4632	0.0870	13.450
BLIINDS-II	0.4585	0.4404	13.487
BRISQUE	0.6597	0.6318	11.405
CORNIA	0.6834	0.6497	11.079
NIQE	0.6066	0.5266	12.065
Proposed	0.7117	0.6035	8.0729

Table 2. Comparison with existing IQA algorithms for CSIQ dataset.

Algorithm	PLCC	SROCC	RMSE
DIIVINE	0.8993	0.8716	0.1253
BLIINDS-II	0.8930	0.8766	0.1290
BRISQUE	0.9275	0.9026	0.1071
CORNIA	-	-	-
NIQE	0.9272	0.8925	0.1090
Proposed	0.7322	0.6651	0.1015

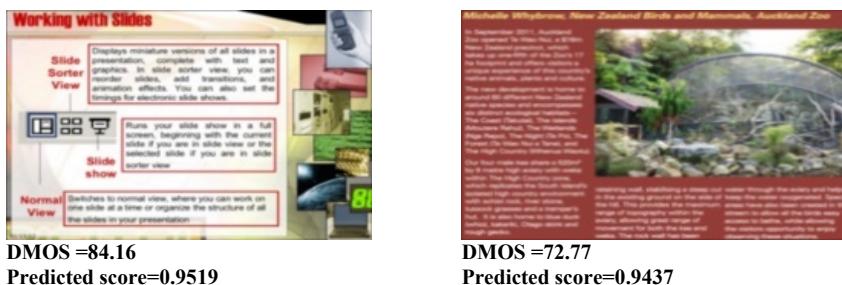


Figure 2. Examples of the results of the proposed method

4. Conclusion

This paper presented a hybrid NR IQA method that enriches traditional NR IQA with deep transfer learning. The method generates a PRI from the input distorted image and then uses pre-trained deep feature extractors to produce feature maps for both images. LSS is used to measure the similarity between the feature maps followed by a nonlinear mapping function to compute the final quality score of the input distorted image. The preliminary results demonstrated that our method had achieved a promising performance. Future work will focus on employing more deep feature extractors to improve the proposed method's performance further.

Acknowledgment

This research was partly supported through Project PID2019-105789RBI00.

References

- [1] Hu J, Wang X, Shao F, Jiang Q. TSPR: Deep network-based blind image quality assessment using two-side pseudo reference images. *Digital Signal Processing*. 2020; 106:102849
- [2] Yang G, Wang Y. Deep Superpixel-based Network for Blind Image Quality Assessment. arXiv preprint arXiv:211006564. 2021
- [3] Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*. 2012;21(12):4695-708
- [4] Moorthy AK, Bovik AC. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*. 2011;20(12):3350-64
- [5] Mittal A, Soundararajan R, Bovik AC. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*. 2012;20(3):209-12
- [6] Li Q, Lin W, Fang Y. No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Processing Letters*. 2016;23(4):541-5
- [7] Wu Q, Wang Z, Li H, editors. A highly efficient method for blind image quality assessment. 2015 IEEE International Conference on Image Processing (ICIP); 2015: IEEE
- [8] Min X, Gu K, Zhai G, Liu J, Yang X, Chen CW. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia*. 2017;20(8):2049-62
- [9] Kang L, Ye P, Li Y, Doermann D, editors. Convolutional neural networks for no-reference image quality assessment. Proceedings of the IEEE conference on computer vision and pattern recognition; 2014
- [10] Saad MA, Bovik AC, Charrier C. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing*. 2012;21(8):3339-52.

- [11] Ye P, Kumar J, Kang L, Doermann D, editors. Unsupervised feature learning framework for no-reference image quality assessment. 2012 IEEE conference on computer vision and pattern recognition; 2012: IEEE
- [12] Larson EC, Chandler DM. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*. 2010;19(1):011006
- [13] Yang H, Fang Y, Lin W. Perceptual quality assessment of screen content images. *IEEE Transactions on Image Processing*. 2015;24(11):4408-21
- [14] Wang Z, Bovik AC. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*. 2006;2(1):1-156

Detecting the Area of Bovine Cumulus Oocyte Complexes Using Deep Learning and Semantic Segmentation

Georgios ATHANASIOU^a, Jesus CERQUIDES^a, Annelies RAES^b, Nima AZARI-DOLATABAD^b, Daniel ANGEL-VELEZ^b, Ann VAN SOOM^b, and Josep-Lluis ARCOS^a

^a*Artificial Intelligence Research Institute (IIIA), CSIC, Campus UAB, 08193 Bellaterra, Spain*

^b*Department of Internal Medicine, Reproduction and Population Medicine, Ghent University, 9820 Merelbeke, Belgium*

Abstract. The cumulus-oocyte complex (COC) is an oocyte surrounded by specialized granulosa cells, called cumulus cells. The cumulus cells surrounding the oocyte ensure healthy oocyte and embryo development. The maturity of COCs at oocyte retrieval may be used as an indicator to predict outcome of assisted reproductive technology (ART). Segmenting COCs is a preliminary step in many image processing pipelines to evaluate maturity. However, acquiring well-annotated bright-field microscopy image datasets remains a time-consuming and inaccurate procedure, for most biological domains. Additionally, specialists often partially disagree on their annotations, not only among each other, but also among their own annotations, leading to an inconsistent outcome. Despite the recent advancements in deep learning and image segmentation tools for biological and biomedical images, there is limited usage of them for having more accurate and automated procedures. In this work, we propose an automated pipeline to segment bovine COCs in bright-field microscopy images. The results of our evaluation show that our pipeline is able to segment COCs with the same level of quality as provided by human experts.

Keywords. Deep Learning, Bright-Field Microscopy, Biomedical Imaging, Image Segmentation, Image Analysis

1. Introduction

Infertility is defined as a failure to achieve clinical pregnancy of 12 months or more of regular, unprotected intercourse, and is a big issue for medicine and society. Once the disease is diagnosed, the treatments involve the techniques of Assisted Reproductive Technology (ART). Methods of ART are considered the intracytoplasmic injection of sperm (ICSI) and the in-vitro fertilization (IVF). These methods require several sub-steps, among of which the characterization of morphological characteristics of oocyte and embryo biology elements.

Cumulus expansion is a key element for characterizing the quality of mammalian oocytes, for later use in in-vitro fertilization (IVF). There are several methods for mea-

suring cumulus expansion described in the literature (Chen et al. [1], Ploutarchou et al. [2]). All the methods available are time-consuming, and depend deeply on human subjectivity since the annotation might vary from one expert to another. Some of the methods for measuring the cumulus expansion rely on assessing the area of cumulus including the oocyte. With the aim to help in the automation of these methods, in this work, we propose a pipeline for segmenting the cumulus oocyte complex (COC), since after segmentation, measuring the size of the COC is a very simple task.

Deep learning and Convolutional Neural Networks (CNN) have seen great progress in the use of medical image segmentation in the recent years, offering a positive impact in medicine and healthcare. Image segmentation is a process of breaking an image into smaller parts, creating a representation more meaningful to be processed by machines. In this work, image segmentation is used to segment bright-field microscopy images of cumulus oocyte complexes in immature and mature oocytes, to later compute the relative cumulus expansion, using a U-Net network architecture.

Literature in image segmentation for oocyte microscopy images is not very broad. Firuzinia et al. [3] applied image segmentation methods on human metaphase II mature oocytes, focusing on several morphological characteristics at this stage, and using a total number of 1009 images. Targosz et al. [4] used image segmentation on human oocytes of different phases (MII, MI, PI, DYS, DEG). A dataset of 334 pictures with one or more oocytes was used. There is no clue for these two approaches that the annotation of the oocytes was performed by more than one specialist. Also, both approaches used already pre-trained networks, such as ResNet and MobileNet, and a variety of data augmentation techniques.

There are other applications of image segmentation for supporting Assisted Reproduction Technology techniques, using bright-field microscopy images. The main focus is on early-stage human embryo development to characterize morphological characteristics. Fukunaga et al. [5] developed a system of automating the detection of pronuclei on 900 embryos. Khan et al. [6] and Leahy et al. [7] applied segmentation techniques for counting the number of cells, while there are works (Dirvanauskas et al. [8], Liu et al. [9], Malmsten et al. [10][11][12], Lau et al. [13], Gingold et al. [14], Meseguer et al. [15]) on identifying the development stage.

To the best of our knowledge, this is the first research on image segmentation in bovine oocytes of bright-field microscopy images. The size of the dataset is just 100 oocytes in total, significantly smaller than any of the already mentioned ones. Last, it is the first approach of trying to measure cumulus expansion, and also exploring the effect of the inconsistency and disagreement among several experts. Our results show that the proposed deep learning model could replace humans in segmenting COCs, and that transfer learning is a key component in the training of our model.

The rest of the paper is organized as follows: in Section 2 we present our segmentation pipeline, with details on the network and the techniques used. In Section 3 we introduce the experiments carried out, with an insight on the dataset followed by the results. Section 4 contains a brief discussion and the conclusions, and we finish the paper with some future perspectives.

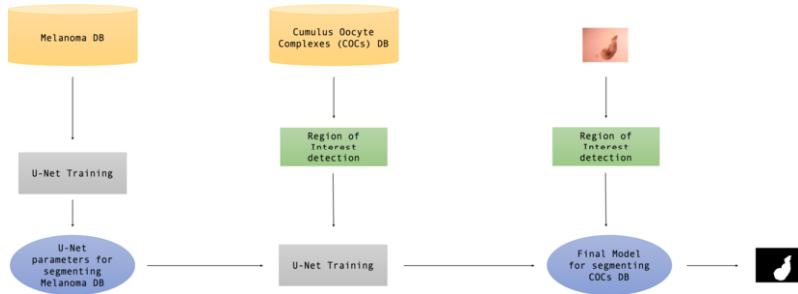


Figure 1. Proposed segmentation training architecture.

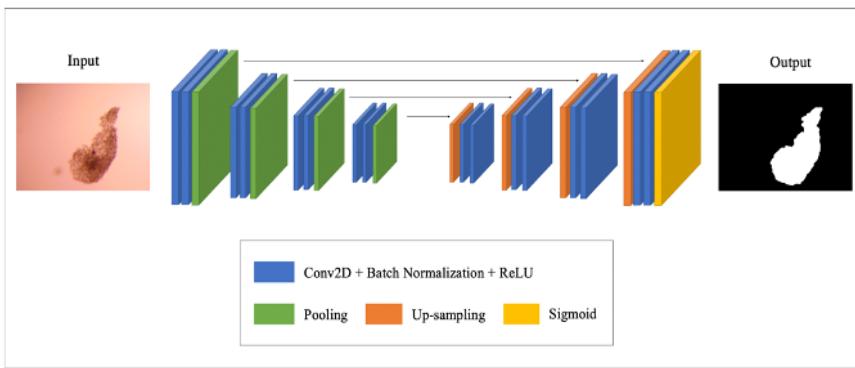


Figure 2. U-Net Architecture

2. Proposal

Our proposed pipeline is presented in Figure 1. The main segmentation model relies on convolutional neural networks (in particular, we rely on the U-Net [16] architecture).

The left hand side of Figure 1 shows our usage of *transfer learning* to overcome the lack of training data from the domain. Specifically, in a first stage, U-Net is pre-trained on a publicly available dataset of a related domain, containing bright-field microscopy images of a melanoma cells.

Furthermore, we split the COC segmentation task in two stages. In the first stage, we use local entropy to perform a very rough segmentation and use it to determine a region of interest (ROI), that is, a bounding box containing the COC. The second stage takes as input the image of the ROI and produces a fine segmentation using the U-net model.

2.1. Network Architecture

A U-Net architecture[16] is adopted for all the experiments. U-Net structure for convolutional neural networks have provided satisfying results in the last years for segmenting biomedical image datasets. In Figure 2, there is a representation of the architecture used for our approach.

The contraction path consists of four blocks of two 3x3 convolutional layers, followed by a ReLU layer and a 2x2 max-pooling layer of stride 2, followed by a same

block with an added dropout layer of $p = 0.5$. The expansive path consists of four blocks of a transposed convolutional layer for up-sampling, a concatenated layer, two 3x3 convolutional layers, a ReLU layer, and afterwards, a last convolutional layer. The proposed architecture was implemented using Keras open-source package and TensorFlow as a back-end platform.

2.2. Loss Functions and Evaluation Metrics

To determine the accuracy of the proposed segmentation we use the Dice Coefficient [17]. Dice coefficient is an indicator of the spatial overlap between two areas, ranging from 0 to 1, with 0 denoting no overlap at all, and 1 denoting perfect overlap. The equation is as follows (1):

$$Dice(f, x, y) = \frac{2 \sum_{ij} f(x)_{ij} y_{ij}}{\sum_{ij} f(x)_{ij} + \sum_{ij} y_{ij}} \quad (1)$$

where y is the ground truth, x is the input image, $f(x)$ is the prediction of the model.

Because Dice Coefficient was considered as the evaluation metric, we selected Dice Loss measure to train the weights of the U-Net architecture. Specifically, Dice Loss function can be expressed as the following equation:

$$loss_{Dice}(f, x, y) = 1 - Dice(f, x, y) \quad (2)$$

2.3. Transfer Learning and Data Augmentation

Transfer learning in machine learning is a technique of using knowledge that has been previously acquired from a model, trained to perform a specific task, to a different but somehow related task. The advantage of using this technique is the reduction of the required data size to train a new model, providing a way of building models without requiring large amounts of data, especially in domains that it is highly difficult to find available data, or the labeling of them is time-consuming. In the current approach, annotating the images takes long, and requires specialists with deep knowledge in the domain, while the required annotations are not available in the first place. For the purposes of this application, an open-source dataset of melanoma images is used¹, and then the models are fine-tuned with a small bunch of images and their corresponding annotations.

Data augmentation in Machine Learning is a set of techniques used to increase the amount of available data by creating modified copies of the given data. Since the available COCs dataset is relatively small, data augmentation is used to increase the randomness of the samples, by flipping images along the axes (horizontally, vertically) or rotating them (90, 180, 270 degrees). In that way, for each iteration, there was some percentage of the given images (and corresponding masks) modified, achieving a better generalization of the approach.

¹<https://challenge.isic-archive.com/data/>

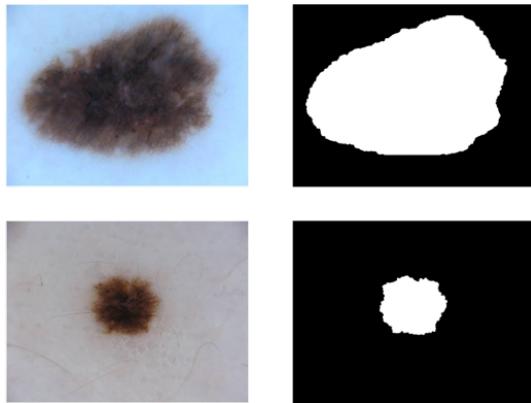


Figure 3. Melanoma dataset sample

3. Experiments

In this section we describe the experimental settings as well as the results of the segmentation pipeline. We start by describing the data used in Section 3.1, then we describe the procedure used for training in Section 3.2, and give the experimental results in Section 3.3.

3.1. Datasets

We have used two different datasets: an already existing dataset for pretraining the segmentation model, and a cumulus oocyte dataset which we have created for the task. We describe each of them next.

3.1.1. Pretraining dataset: Melanoma

The dataset was retrieved from the ISIC 2017 Challenge dataset for Skin Lesion Analysis for melanoma detection [18]. It contains 2.000 RGB images manually segmented by medical specialists and forming binary masks for each image (Figure 3). The images and the masks are translated to greyscale and rescaled to 192x240 pixels, before being fit to the CNN, to match its input dimensions.

3.1.2. Cumulus Oocyte Complexes (COCs) dataset

We have created a dataset of bovine cumulus oocyte complexes. It contains images from 100 oocytes. The COCs were incubated for 22 hours, at 38.5 °C, in 5% CO₂ in humidified air [19]. They cultured in tissue culture medium (TCM)-199, supplemented with epidermal growth factor (EGF) and gentamicin, while each oocyte was individually matured in 20µL droplets; briefly, 17 droplets of 20µL medium each were prepared in Petri dishes (60 × 15 mm; Thermo Fisher Scientific, Waltham, MA USA) and covered with 7.5 mL paraffin oil. The microscope used for the pictures was Olympus stereomicroscope, at 56x magnification, using a TOUPCAM UCMOS05100KPA camera and the ImageJ software. The initial size of the images taken were 1944x2592 pixels, and they are all

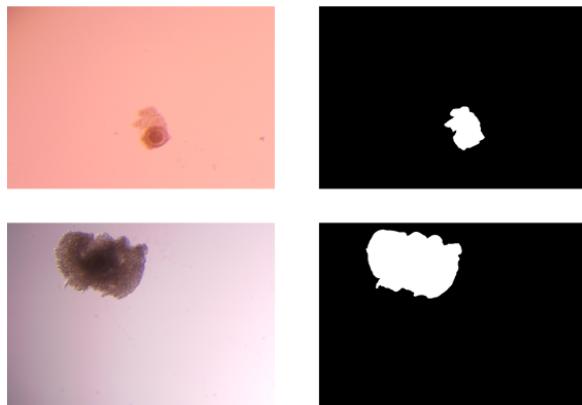


Figure 4. Cumulus Oocyte Complex (COC) dataset sample

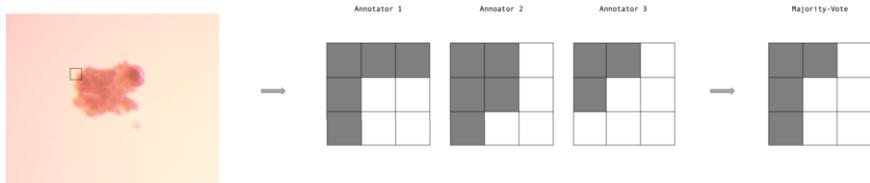


Figure 5. Majority-vote idea.

saved as png. For each of the oocytes we have an immature image (before incubation) and a mature image (after incubation).

We requested three specialists [A1, A2, A3] to segment the images using the ImageJ software, and saving the masks as .png of the same size of the original images. Since the masks provided by each of the three annotators were slightly different we created a consensus segmentation from the three masks. Each pixel of the consensus segmentation was marked as being part of the COC if at least two annotators have marked it as part of the COC in their respective segmentations. For example, if a pixel has 2 positive votes of being part of the COC, and 1 negative, then it is considered as part of the COC. Similarly, we proceed for all of them. The idea is presented in Figure 5 for a random example, where for a 9x9 pixel-square, the majority vote for every pixel is translated to the final output.

For training the model, the previously decided masks were used as ground truth, and from now on, this dataset will be referred to as the majority-vote dataset. The images and the masks were translated to greyscale and to the same ratio as the melanoma dataset, at 192x240, using OpenCV's area interpolation. For the final evaluation, the size of the masks was set back to the initial size (1944x2592), using OpenCV's cubic interpolation. The final results and conclusions remained unaffected by the rescaling process. Other alternatives of combining the information of the experts were considered, such as having probability pixels, instead of deterministic ones, but they were left to be studied in future work.

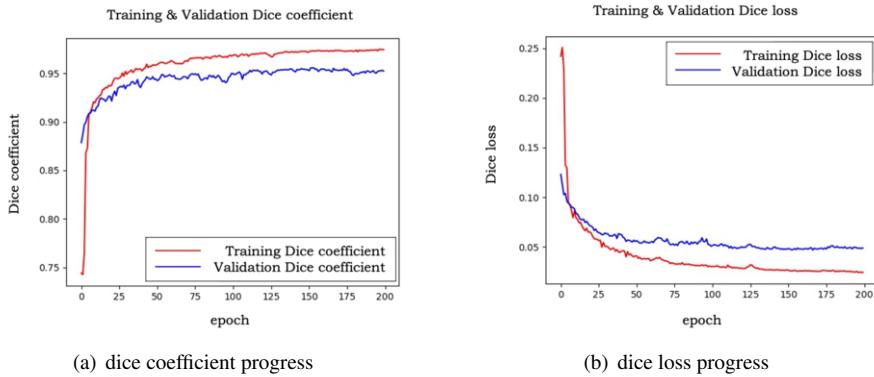


Figure 6. Evolution of mean dice coefficient and mean dice loss.

3.2. Experimental procedure

To evaluate the segmentation models we use 10-fold cross-validation, since the dataset size is limited, and this approach provide safe results. At each fold, 90 oocytes (180 images) were used for training and 10 oocytes (20 images) for validation. The model was trained on minibatches of 32 for 200 epochs. For each fold, we generated the masks for the 20 validation images, resulting in a total of 200 masks after going through all the folds. Then, we compared the masks thus generated, to the masks provided by each of the experts. To evaluate the similarity between two masks we use the dice coefficient. Figure 6 shows the evolution of the mean of dice coefficient and of the mean of the dice loss across all ten folds. The mean dice performance of the majority-vote model is high converging at around 95%.

We are interested in comparing our method against human annotators. To do that we compute the similarity among the annotations of each of the three experts and compare them with the similarity between the each of the human experts and our proposed method. Since dice values do not follow a normal distribution, we use the median of the dice coefficient to evaluate the similarity between any two annotators. This median is expected to be 100% for a perfect similarity and 0% for completely disagreeing annotators.

3.3. Results

The first three rows of Table 1 show how similar are the segmentations of the COC between each pair of human annotators. We see that the numbers are in the 95.15%-95.63% range. In our case, the deep learning model proposed reaches a range between 95.99%-96.48%, higher than the one among the experts. This allows us to consider the results of our model indistinguishable from those of a different human expert. Taking into account the cost of annotating, our method should be considered as a very reasonable alternative to annotation by means of human experts.

To understand the value of each of the components in our pipeline (namely, ROI focusing and transfer learning) we have run an ablation study, removing each of them. The first one concerns the same model, without using the pre-processing stage for detecting the region of interest (ROI), but keeping the transfer learning approach from the

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	-	95.15%	95.49%
Annotator 2	95.15%	-	95.63%
Annotator 3	95.49%	95.63%	-
Our proposal	96.32%	95.99%	96.48%
without ROI	95.95%	95.61%	95.97%
without RT	13.17%	13.55%	14.00%
without TL	no convergence		

Table 1. Comparison of the median of dice coefficients of different models with the ones of human specialists.

melanoma dataset. The range reached from this configuration (*without ROI* in Table 1) is still very reasonable, being in the same levels of experts' performance (95.61%-95.97%); just a bit lower than the final proposal. On the other hand, when we use the model trained only on the melanoma dataset, without any further retraining (RT) on the COC dataset, the performance of the model only reached values in the range of 13.17%-14.00% (*without RT*). Finally, when we remove transfer learning from our proposal, just starting from random initial weights instead (*without TL* in Table 1), the model failed to converge, concluding that transfer learning is essential for this task, possibly due to our limited access to annotated images.

It is also interesting to provide a visual representation of the region of interest of the segmented images, to better comprehend what is going on the segmented areas (Figure 7). Figure 7(a) shows a cumulus oocyte. In the next row (Figures 7(b), 7(c), 7(d)), show the masks, as were provided by the experts themselves. It is pretty obvious that they are not really coincident, especially around the borders of the oocyte, while it is clear that some of them are more detailed on annotating the perimeter, while others propose a more smooth perimeter. These differences affect the way a model is able to be trained, with a ground truth being controversial. Below, Figure 7(e) shows the mask generated using the majority vote of the three masks above. After that, Figure 7(f), presents the mask generated by using the proposed model. Visually, it is almost identical to the one in Figure 7(e), noting also that the perimeter is smoother than annotator 1's approach, but it tries to keep some important details.

4. Conclusions

In the recent years, there is an increase use of deep learning and image segmentation techniques in Assisted Reproductive Technology field. Some attempts have been made in identifying morphological characteristics from oocyte and embryo bright-field microscopy images, the majority of them in for human species. However, there is limited use in other mammalian species, and no use at all for segmenting bovine cumulus oocyte complexes.

The current research is focusing on segmenting the COCs out of a small-sized dataset. This approach presented a supervised method of detecting the cumulus, using transfer learning of a related domain of melanoma images. The reported dice coefficient of the models proved that the best performing model, using majority-vote annotations for training, is promising, since the scores are identical to the human ones.

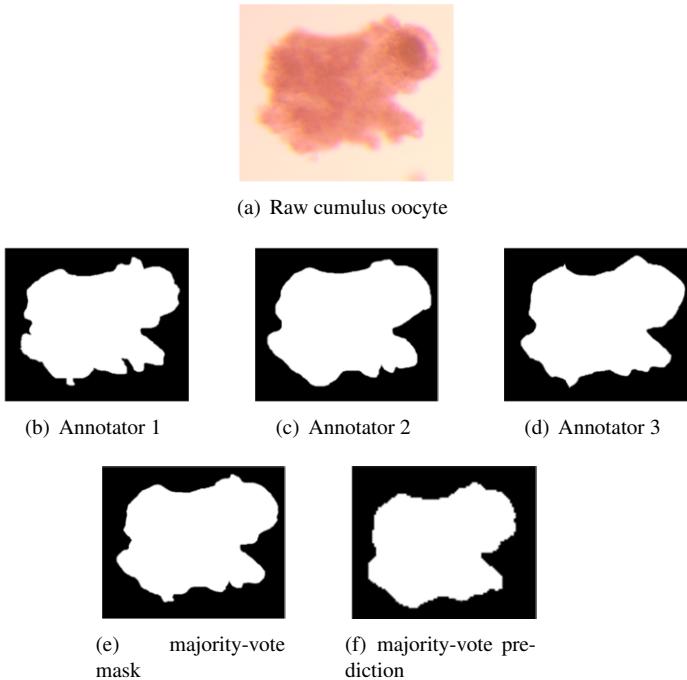


Figure 7. Visual comparison of experts masks and models predictions.

Examining the conditions of the problem in-depth, it became clear that the experts slightly differ when they are annotating or evaluating cumulus oocytes image datasets. Some of them are very detailed and try to be as accurate as possible (Figure 7(b)), without considering the time cost. Others, are less detailed and focus on providing the results faster, leading to smoother, perhaps not so accurate annotations (Figure 7(c), Figure 7(d)). However, trying to find the most beneficial approach, it is rather puzzling to decide and weight more on one of them.

The proposed method of using a majority-vote model, a model that decides if a pixel is part of the cumulus oocyte depending on what the majority of the experts indicates, intents to tackle the issue of partial disagreement among several annotators.

According to the median dice coefficient results, the proposed deep learning model outperforms the human performance, as it is mentioned and presented in Table 1. Noticeably, even with a small-sized dataset and inconsistency among experts of what should be considered as cumulus oocyte part, deep learning algorithms exhibit high and consistent performance, offering more accurate results and a time-saving method.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860960 and by project CI-SUSTAIN (PID2019-104156GB-I00) funded by the Spanish Ministry of Science and Innovation. Georgios Athanasiou is a PhD Student of the doctoral program in Computer Science at the Universitat Autònoma de Barcelona.

References

- [1] Chen L, Russell PT, Larsen WJ. Functional significance of cumulus expansion in the mouse: Roles for the preovulatory synthesis of hyaluronic acid within the cumulus mass. *Molecular Reproduction and Development*. 1993;34(1):87-93.
- [2] Ploutarchou P, Melo P, Day A, Milner C, Williams S. Molecular analysis of the cumulus matrix: Insights from mice with O-glycan-deficient oocytes. *Reproduction*. 2015 May;149:533-43.
- [3] Firuzinia S, Afzali SM, Ghasemian F, Mirroshandel SA. A robust deep learning-based multiclass segmentation method for analyzing human metaphase II oocyte images. *Computer Methods and Programs in Biomedicine*. 2021 Apr;201:105946.
- [4] Targosz A, Przystalka P, Wiaderkiewicz R, Mrugacz G. Semantic segmentation of human oocyte images using deep neural networks. *BioMedical Engineering OnLine*. 2021 Apr;20(1):40.
- [5] Fukunaga N, Sanami S, Kitasaka H, Tsuzuki Y, Watanabe H, Kida Y, et al. Development of an automated two pronuclei detection system on time-lapse embryo images using deep learning techniques. *Reproductive Medicine and Biology*. 2020;19(3):286-94.
- [6] Khan A, Gould S, Salzmann M. Segmentation of developing human embryo in time-lapse microscopy. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016. p. 930-4. ISSN: 1945-8452.
- [7] Leahy BD, Jang WD, Yang HY, Struyven R, Wei D, Sun Z, et al. Automated Measurements of Key Morphological Features of Human Embryos for IVF. *arXiv:200600067 [cs, q-bio]*. 2020 Jul. ArXiv: 2006.00067.
- [8] Dirvanauskas D, Maskeliunas R, Raudonis V, Damasevicius R. Embryo development stage prediction algorithm for automated time lapse incubators. *Computer Methods and Programs in Biomedicine*. 2019 Aug;177:161-74.
- [9] Liu Z, Huang B, Cui Y, Xu Y, Zhang B, Zhu L, et al. Multi-Task Deep Learning With Dynamic Programming for Embryo Early Development Stage Classification From Time-Lapse Videos. *IEEE Access*. 2019;7:122153-63. Conference Name: IEEE Access.
- [10] Malmsten J, Zaninovic N, Zhan Q, Toschi M, Rosenwaks Z, Shan J. Automatic prediction of embryo cell stages using artificial intelligence convolutional neural network. *Fertility and Sterility*. 2018 Sep;110(4):e360. Publisher: Elsevier.
- [11] Malmsten J, Zaninovic N, Zhan Q, Rosenwaks Z, Shan J. Automated cell stage predictions in early mouse and human embryos using convolutional neural networks. In: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI); 2019. p. 1-4. ISSN: 2641-3604.
- [12] Malmsten J, Zaninovic N, Zhan Q, Rosenwaks Z, Shan J. Automated cell division classification in early mouse and human embryos using convolutional neural networks. *Neural Computing and Applications*. 2020 Jun.
- [13] Lau T, Ng N, Gingold J, Desai N, McAuley J, Lipton ZC. Embryo staging with weakly-supervised region selection and dynamically-decoded predictions. *arXiv:190404419 [cs]*. 2019 Apr. ArXiv: 1904.04419.
- [14] Gingold JA, Ng NH, McAuley J, Lipton Z, Desai N. Predicting embryo morphokinetic annotations from time-lapse videos using convolutional neural networks. *Fertility and Sterility*. 2018 Sep;110(4):e220. Publisher: Elsevier.
- [15] Meseguer M, Herrero J, Tejera A, Hilligsoe KM, Ramsing NB, Remohí J. The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction (Oxford, England)*. 2011 Oct;26(10):2658-71.
- [16] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:150504597 [cs]*. 2015 May. ArXiv: 1505.04597.
- [17] Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302.
- [18] Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv:160501397 [cs]*. 2016 May. ArXiv: 1605.01397.
- [19] Azari-Dolatabad N, Raes A, Pavani KC, Asaadi A, Angel-Velez D, Van Damme P, et al. Follicular fluid during individual oocyte maturation enhances cumulus expansion and improves embryo development and quality in a dose-specific manner. *Theriogenology*. 2021 May;166:38-45.

Object Segmentation of Cluttered Airborne LiDAR Point Clouds

Mariona CARÓS^a, Ariadna JUST^b Santi SEGUÍ^a and Jordi VITRIÀ^a

^aDepartament de Matemàtiques i Informàtica, Universitat de Barcelona (UB),

Gran Via Corts Catalanes, 585, 08007 Barcelona, Spain

^bInstitut Cartogràfic i Geològic de Catalunya, Barcelona, Spain

Abstract. Airborne topographic LiDAR is an active remote sensing technology that emits near-infrared light to map objects on the Earth's surface. Derived products of LiDAR are suitable to service a wide range of applications because of their rich three-dimensional spatial information and their capacity to obtain multiple returns. However, processing point cloud data still requires a large effort in manual editing. Certain human-made objects are difficult to detect because of their variety of shapes, irregularly-distributed point clouds, and a low number of class samples. In this work, we propose an end-to-end deep learning framework to automatize the detection and segmentation of objects defined by an arbitrary number of LiDAR points surrounded by clutter. Our method is based on a light version of PointNet that achieves good performance on both object recognition and segmentation tasks. The results are tested against manually delineated power transmission towers and show promising accuracy.

Keywords. LiDAR, Point Clouds, Deep Learning, Segmentation, Remote Sensing

1. Introduction

Light Detection and Ranging (LiDAR) is a technology that emits pulses of light to measure the distance from the sensor to the objects. Topographic LiDAR sensors generate pulses of near-infrared light that are reflected on the Earth's surface to create high-resolution three-dimensional (3D) maps of the surrounding environment. The data can be used for a variety of purposes, such as environmental monitoring [1], forest inventories [2] or object detection [5].

Laser scanning can be mounted on different platforms depending on the target purpose. For indoor mapping or architecture heritage, Terrestrial Laser Scanning (TLS) is used, which is commonly placed on tripods. For road and 3D urban mapping, Mobile Laser Scanning (MLS) is the choice, usually mounted on vehicles. For large-scale applications including terrain modeling, forestry, and urban mapping, an Airborne Laser Scanning (ALS) is generally used; and for global monitoring of terrain and vegetation, LiDAR is installed in a satellite (SLS).

LiDAR observations are stored as point clouds, which are collections of points defined by 3D coordinates and may include features like intensity, number of return, or incidence angle. Recent airborne LiDAR systems are hybrid in the sense they include an RGB and Near-Infrared (NIR) camera, capturing simultaneous images and assigning

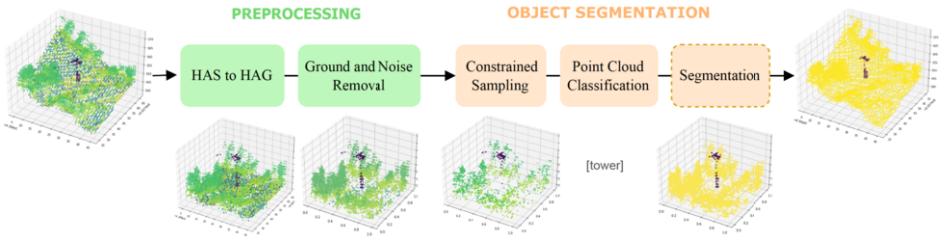


Figure 1. Overview of our preprocessing and segmentation pipelines. Point heights are transformed from heights above sea (HAS) into heights above ground (HAG). Then, the point cloud is filtered from noise, constrained sampling is applied, and the resulting point cloud is used to train both the classifier and segmentation networks. During inference, the classification network detects if the object is within the processing point cloud, if it does, segmentation is applied to all input points. Otherwise, the block is skipped and all points are labeled as background.

4 channels of color to each LiDAR point cloud. The advancements of this 3D sensing technology are increasing the demand for point cloud processing techniques. Simultaneously, given the highly-accurate 3D information, point cloud classification and segmentation have become an active research direction in the fields of remote sensing and computer science. There have been remarkable advances in deep learning techniques for point cloud understanding such as PointNet [7], PointNet++[8], RSNet [11], Point Graphs [10], GSPN [9] or Point Transformer [12], and many of these approaches achieve impressive results. However, almost all of them are limited to synthetic data [4] or indoor datasets [3] and are difficult to be directly extended to real large-scale outdoor airborne LiDAR data, where the point clouds have significantly different number and distribution of points. When dealing with automated 3D mapping with airborne LiDAR data, certain human-made objects are difficult to detect because of their irregularly-distributed point clouds and variety of shapes. Partial observations of objects are common due to occlusions, and background is blended with objects due to clutter in the real-world scenes. In addition, the low number of class samples makes it difficult to use deep learning models.

In order to overcome these challenges, we present an end-to-end framework for object recognition and segmentation of LiDAR data and describe a real application case study. The proposed approach consists of a preprocessing stage with a sliding window to split our point cloud data into cubes, and a deep learning framework based on PointNet [7] to detect if the object is within the cube and segment it. A general overview of the framework is shown in Figure 1.

Specifically, in this work we make the following main contributions. First, we describe an end-to-end approach for processing large-scale airborne LiDAR data with deep learning. Second, we propose a 3D detection framework capable of segmenting a specific object defined by an arbitrary number of points. Third, we report several experiments on our manually delineated airborne LiDAR dataset containing power transmission towers, where we conduct controlled studies to examine specific choices and parameters. Our code is publicly available at <https://github.com/marionacaros/3d-object-segmentation>.

This article is structured as follows: Section 2 overviews literature and related work in regards to 3D data modeling. Section 3 presents the proposed method. The dataset is explained in section 4. In section 5 we report experiments and results. Finally, section 6 summarizes the main conclusions.

2. Related Work

Deep learning on 3D data has been receiving increasing attention in recent years. A number of different representations have been explored, including multi-view projections, voxel grids, and point clouds.

LiDAR data are irregularly distributed, unordered and scattered. Considering the success of Convolutional Neural Networks (CNNs) for image understanding, an intuitive approach is to project volumetric data into 2D planes [14]. Then, CNNs are used to extract feature representations from each of the planes. Nevertheless, these multi-view projections collapse spatial information of points which may affect object recognition. An alternative approach to use the benefits of CNNs on LiDAR data is to convert raw point cloud data into voxelized grids [6]. The irregular distribution of data is transformed into a uniform grid where 3D CNNs are applied. Compared to multi-view methods, this strategy has no loss of information, but can be very expensive in terms of computational and memory costs due to the generated number of voxels.

Rather than projecting irregular point clouds onto regular grids, point-based networks directly process point clouds. PointNet [7] was the pioneer work to directly process point sets. The key idea of PointNet is to process points independently by using permutation-invariant operators, and then aggregate them into a global feature representation by max-pooling. In the following work of these authors, PointNet++[8], they incorporate local dependencies and hierarchical feature learning in the network to increase sensitivity to local geometric layout. A number of works relate point clouds to graphs and perform graph convolutions for feature extraction [10]. RSNet [11] uses Recurrent Neural Network (RNN) and a slice unpooling layer to project features of unordered points onto an ordered sequence of feature vectors. Point Transformer [12] exploits the positional information of points and applies self-attention to point clouds for semantic scene segmentation and object classification.

In this work, we select PointNet architecture as the base of our framework due to its simplicity, robustness and low execution time [13]. The goal of our work is to provide an end-to-end framework for real-world LiDAR applications. Thus, we not only focus on accuracy but memory consumption and training time as well.

3. Proposed Approach

Given a point cloud and a candidate label set, our task is to assign each of input points with one of the semantic labels in order to identify a specific object. Our method consists of two main stages presented in Figure 1: Preprocessing and Object Segmentation. Preprocessing blocks make a transformation to the data to simplify the segmentation task.

3.1. Preprocessing

The first step of our algorithm, in both train and inference tasks, is to partition the point cloud into smaller cubes, which enables parallelization and makes the segmentation task easier for the model. At the end of the pipeline prediction of cubes are merged again in the same point cloud. The size of the cubes is chosen considering two factors: The size of the object to be detected, and the minimum distance between objects.

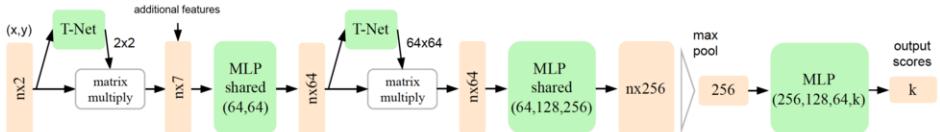


Figure 2. The classification network follows the PointNet structure. It takes (x, y) values of n points as input, applies a transformation learned by T-Net, and it concatenates additional features to the output; such as z , intensity, and RGB. Then, another transformation is applied to the feature space. Finally, features are aggregated by max pooling. The output is classification scores for k classes. MLP stands for multi-layer perceptron.

Once we have our reduced point clouds, we transform heights above sea (HAS) into heights above ground (HAG). Shapes may not be easily compared in hilly or mountainous terrain, because part of the observed variability is due to changes in the altitude of the surface. For this reason, we normalize all heights by subtracting ground's height¹ to z coordinate values and obtain z_{HAG} . Next, we remove ground points ($z_{HAG} = 0$) and noise (i.e. points above 100 meters). We consider that these types of points provide no valuable information to identify our target object, and by discarding them we increase the efficiency and performance of our model. Finally, we normalize them into a unit sphere.

3.2. Object Segmentation

The goal of this stage is to segment the target object within the point cloud window. The methodology is the following; The input point cloud is sampled to a fixed number of points, then a binary classifier is used to detect if the object is within the point cloud. If the result is negative, all points are labeled as background and the subsequent window is processed. On the contrary, the whole point cloud is fed into the segmentation model, which assigns a label to each point.

The algorithm behind both of our models is a lighter version of PointNet, where the number of parameters is reduced to one quarter. The classification network is shown in Figure 2. We input x and y coordinates into the transformation net (T-Net) which learns a canonical representation of these 2 dimensions. Then, z and additional features such as intensity, RGB, and NIR are concatenated. Notice that z is not input into the canonical transformation because we do not expect our target objects to be tilted. Finally, features are aggregated by max pooling. The output of the network is the classification score, we set $k = 2$ for object detection.

Training PointNet requires all input point clouds to be the same size. However, in LiDAR data the number of points can variate by tens of thousands. A usual approach is to randomly sample points from the point cloud and input them into the model in batches of a fixed size. A major problem of this procedure is the sampling of high-dense point sets, where we see a drop in performance caused by incomplete objects. This occurs when objects are small in comparison with clutter and random sampling causes points of the same object to be scattered among batches. To alleviate the aforementioned problem, we propose the **Constrained Sampling** block, which takes advantage of the spatial location of points. We know that most of our dataset points correspond to vegetation, so we design a constrained sampling based on heights. In Figure 3 we present three heights distributions obtained from point cloud windows containing our target object (a power transmission

¹Topographic ground map is provided by an external source [18]

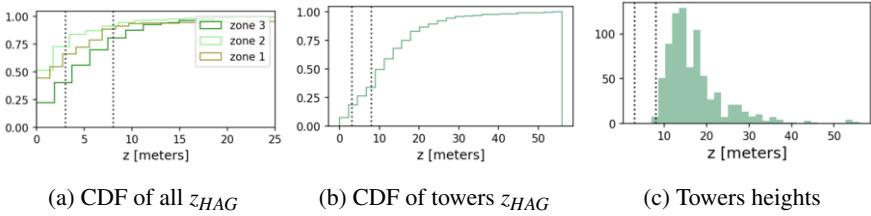


Figure 3. Cumulative distribution function (CDF) of points heights. Dotted lines identify the values of 3 and 8 meters, which correspond to low and medium vegetation in Catalonia.

tower). From left to right we see the cumulative distribution function (CDF) of all z_{HAG} values, z_{HAG} of points labeled as a tower, and maximum z_{HAG} per tower. By comparing Figures 3a and 3b we notice a big difference between distributions of points. Regarding all z_{HAG} values, there are between 40% – 75% of the points in the range of 3 – 8 meters, while the percentage is reduced to half if we focus on points labeled as tower. These values make total sense as 3 and 8 meters correspond to low and medium vegetation in Catalonia, while the minimum height of our target object is 10 meters (shortest tower in Figure 6). In consideration of this, our sampling methodology is the following; We first down-sample points below 3 meters, if the number of points is larger than the defined fixed amount, we down-sample again raising the threshold to 8 meters. Finally, if we still have too many points, we randomly sample the whole point cloud. The goal of our sample strategy is to remove cluttered points and make our target object more visible to the model.

We apply constrained sampling as a previous step to both of our models in the training phase, so all input data is sampled according to this strategy. We compare the obtained results with random sampling in Section 5 and show that applying constrained sampling results in high precision results. During inference, all input points are fed into the segmentation model.

4. Dataset

Our dataset is composed of three large-scale outdoor areas from several ALS flights in Catalonia, each covering 3, 112 and 108 squared kilometers (total of 223 squared kilometers). These data were collected with two different LiDAR sensors by ICGC². The first one is a Leica ALS50-II, which only allows capturing signal intensity data.

²Institut Cartogràfic i Geològic de Catalunya

Table 1. Dataset properties

	Zone 1	Zone 2	Zone 3
Sensor properties	Intensity	Intensity, RGB, NIR	Intensity, RGB, NIR
Mean density [pts/m²]	6	10	8
Blocks size [Km]	0.5×0.5	1×1	2×2
Total number of blocks	12	112	27
Point cloud windows containing towers	18	296	475
Point cloud windows without towers	2,044	53,756	38,296

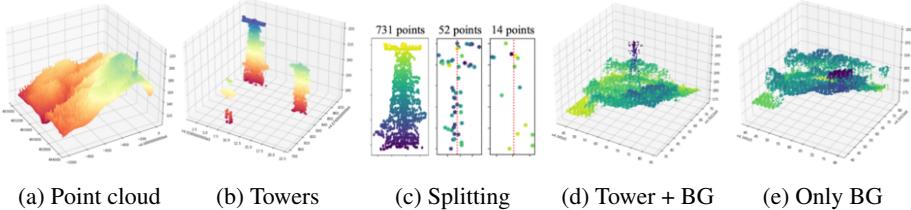


Figure 4. Dataset processing pipeline for training data. From left to right: (a) Input point cloud cube of 1×1 Km. (b) Points labeled as tower. (c) Separate tower point sets. (d) Tower with background points. (e) Only background points.

The second one is a Terrain Mapper 2, which combines a LiDAR sensor with two nadir cameras in RGB and NIR. Each point is defined by xyz coordinates, intensity, RGB, and NIR. The obtained mean point density varies with each flight and sensor, from 6 to 10 pts/m^2 , as it is shown in Table 1.

Points are labeled into four classes: Power transmission tower, power lines, ground, and background. The latter includes all the points different from the previous classes. In order to train a model for power tower recognition, we need each tower in a separate point cloud. These point cloud cubes are obtained by using a sliding window approach that we detail next.

Given a LiDAR point cloud (Figure 4a), we identify points labeled as a power transmission tower and apply a fixed-size sliding window to split them on the xy plane resulting in separate tower point sets. We use a window of 20×20 meters to avoid having two towers in the same point set, as the maximum tower's width in our dataset is 20 meters. Even though in some cases we find narrow towers sharing a window, these cases are very rare. Once towers are segmented (Figure 4c), we store those with more than 19 points, which represents 80% of the towers. The number of points per tower is very variable, as presented in Figure 5, ranging from 1 to 2434, and towers comprising less than 20 points do not present a clear shape. Some examples are illustrated in Figure 4c. Next, we get the center of each tower by computing the mean of (x, y) coordinates. This location is used to get all points within the window, including cables and background. Finally, we normalize points in the unit sphere and store two versions of the same cube; the first one containing all types of points, and a second one with only ground and background points. This is done to facilitate the task of distinguishing between point clouds that have the target object and those that do not.

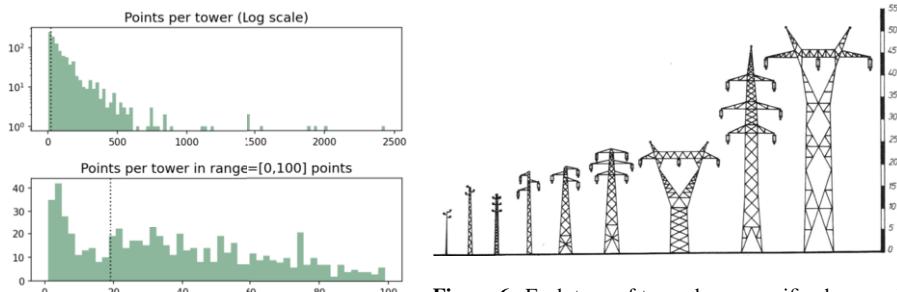
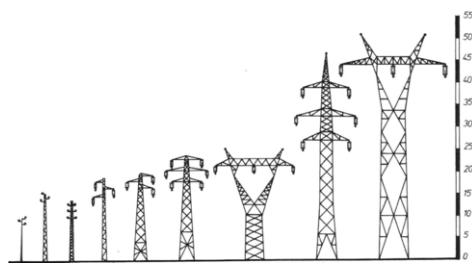


Figure 5. Distribution of points per tower

Figure 6. Each type of tower has a specific shape and height. The highest tower is 50 meters and the shortest is 10 meters.



We end up with a total of 789 point cloud windows containing power transmission towers (see Table 1), which represents only 0.84% of the data. The dataset is highly imbalanced and considering that power transmission towers show a variety of shapes, as presented in Figure 6, it is a challenging task to develop an accurate model for object recognition. Models trained on such data perform poorly for weakly represented classes [16]. However, by using appropriate preprocessing techniques and data augmentation, we achieve good results.

5. Experimentation and Results

In this section, we report the performance of our classification and segmentation models in terms of accuracy and computational cost. We also study the effects of various processing and architecture choices with an ablation study.

5.1. Implementation Details

To address the problem of under-represented positive samples, we use two class balancing strategies: Data augmentation and a class-balancing weighted loss.

We augment the training set by storing a different xy plane of each positive sample, which doubles the number of our target object samples. Specifically, we add a random variation of 10 meters to each x and y to place towers at any part of the xy plane within the point cloud. We use windows of 40 x 40 meters so large towers are not cut in half.

Regarding class-balancing, we weight the negative log-likelihood loss by using the effective number of samples presented in [15]. The effective number of samples is defined as the volume of samples and can be calculated by the following formula $(1 - \beta^n)/(1 - \beta)$, where n is the number of samples and $\beta \in [0, 1]$ is a hyperparameter. The hyperparameter β smoothly adjusts the class-balanced term between no re-weighting and re-weighing by inverse class frequency. We experiment with different β values (0.9, 0.999, 0.9999), as the authors suggest, and compare their results.

We implement our code in Pytorch v1.8. with CUDA 11.6. We use Adam optimizer and an initial learning rate of 0.001 with 0.5 decay when loss does not decrease for 10 epochs. Models are trained for 50 epochs with an early stopping on the validation loss. The system used for the experiments has the following configuration: (i) CPU: Intel Xeon Silver 4210, (ii) RAM: 126GB, (iii) GPU: Quadro RTX 5000 - 16 GB, and (iv) OS: Ubuntu 20.04.

5.2. Experimental Set-Up

Each point is represented by a 7-dim vector of xyz coordinates, intensity, green, blue, and Normalized Difference Vegetation Index (NDVI). NDVI is used in remote sensing [17] to indicate whether or not the target being observed contains live green vegetation, and it is computed by using a simple formula: $(NIR - red)/(NIR + red)$.

The train-test split is done considering the number of blocks with towers. We set aside 10% of blocks with towers for evaluation, which is reported in Table 2, so that during training none of the point clouds belonging to the test blocks are observed. In classification, we use a training set of 80757 point clouds, a validation set of 6970, and a test set of 7935. During segmentation, background samples are reduced to 5%, as they

Table 2. Dataset train-test split

	Zone 1	Zone 2	Zone 3
Blocks with towers	7	70	21
Blocks for train	11	105	24
Blocks for test	1	7	3

do not add value for learning the segmentation task. In training time we use 2048 points and a batch size of 32, while in inference we use all points in batches of 1. For evaluation metrics, we use F₁ score in classification and mean classwise intersection over union (mIoU) in segmentation.

5.3. Quantitative and Qualitative Results

The results of classification and segmentation are presented in Tables 3 and 4, respectively. In both models, the best results are achieved when using all available LiDAR features (intensity, RGB and NIR), together with constrained sampling and weighted loss with β set to 0.999, which yields weights of [0.42, 0.58] in Classification and [0.4, 0.6] in Segmentation. The difference between weights is explained by the fact that in classification positives are defined by point clouds containing a tower, while in segmentation the positives are the points labeled as a tower.

We qualitatively evaluate the segmentation capability of our method to unseen point clouds. Figure 7 shows the best and worst predictions using both uniform and constrained sampling. We can see that the predictions using constrained sampling are very close to the ground truth. Our model captures the whole tower shape when there is vegetation around, and in some cases, it is even able to detect two towers sharing the same point cloud window. When the object is not entirely in the cube some points are missed. Nevertheless, the majority of them are well predicted. The worst predictions are caused by small and low-density towers surrounded by high vegetation, as expected.

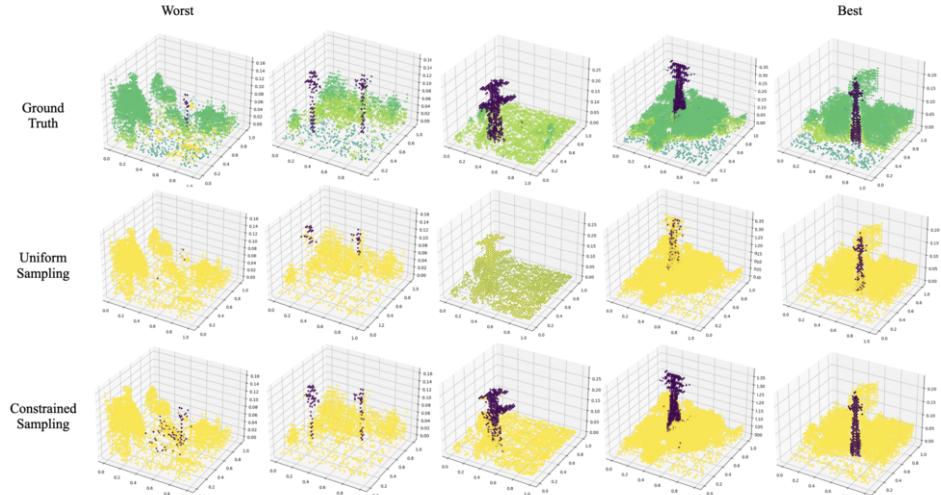


Figure 7. Visualization for object segmentation results. Top row is the ground truth. Bottom and middle rows are results produced by our model with and without constrained sampling.

Table 3. Results of classification and ablation study (features, sampling, and weights). Subscript L denotes light version of PointNet. Sampling* refers to constrained sampling.

Model	RGB+NIR	Sampling*	β	Weights	Time [h]	Mem. [GiB]	PR F1
PointNet $_L$		x	0.999	[0.42, 0.58]	1.6	2.25	84.4
PointNet $_L$	x		0.999	[0.42, 0.58]	2.1	2.25	90.7
PointNet $_L$	x	x	0.9	[0.5, 0.5]	1.5	2.25	89.7
PointNet $_L$	x	x	0.9999	[0.1, 0.9]	3.0	2.25	78.4
PointNet	x	x	0.999	[0.42, 0.58]	3.4	6.7	88.3
PointNet $_L$	x	x	0.999	[0.42, 0.58]	3.0	2.25	92.7

5.4. Ablation Study

We now study the variation in performance involved in each specific decision. In particular, several key features and parameters are considered: (i) The use of RGB and NIR features attached to each point; (ii) the use of constrained sampling; (iii) the effect of class-balancing weights; and (iv) the amount of trainable parameters. Results are shown in Tables 3 and 4.

RGB and NIR. We first investigate the effect of using color and near-infrared information. We can see that without using these extra features the performance of both models drops significantly, being 8.3 absolute percentage points in classification and 22.8 in segmentation. This suggests that RGB and NIR are essential for a proper object segmentation.

Constrained sampling. We study the type of sampling applied to the point clouds. When using random sampling both models perform worst, the models may not be able to learn the shape of the object because of all the clutter in the scene. The use of constrained sampling mainly improves the segmentation model with an increase of 10.5 IoU tower and 0.9 mean IoU. The improvement in object segmentation is visually noticed in Figure 7.

Class-balancing weights. We conduct an ablation study of β . When β is set to 0.9, both classes are assigned the same weight and the model miss to detect a lot of positives. When β is set to 0.9999, the predictions result in a lot of false positives. The best performance is achieved when β is set to 0.99.

Trainable parameters. We compare the performance between the original PointNet and our lighter version (PointNet $_L$), where the number of parameters is reduced from 3.5M to 0.9M. We see that PointNet $_L$ is not only more efficient but outperforms PointNet in terms of precision. This is probably caused by having too many parameters to learn.

Table 4. Results of segmentation and ablation study (features, sampling, and weights). Subscript L denotes light version of PointNet. Sampling* refers to constrained sampling.

Model	RGB+NIR	Sampling*	Weights	IoU tower	IoU veg	mIoU
PointNet $_L$		x	[0.4, 0.6]	28.0	98.0	63.0
PointNet $_L$	x		[0.5, 0.5]	32.9	98.4	65.7
PointNet $_L$	x	x	[0.5, 0.5]	55.0	98.7	76.9
PointNet $_L$	x		[0.4, 0.6]	63.0	98.8	80.9
PointNet	x	x	[0.4, 0.6]	71.7	96.6	84.2
PointNet $_L$	x	x	[0.4, 0.6]	73.5	98.0	85.8

6. Conclusions

We proposed an object segmentation method for point cloud data that is going to be operationally implemented in ICGC's production lines. Our approach enables the segmentation of objects defined by variable point sets given a large outdoor scene. We report several experiments on our manually delineated airborne LiDAR dataset, where we conduct controlled studies to examine specific choices and parameters. We conclude that color and near-infrared features are essential for proper object segmentation, constrained sampling is key in cluttered point clouds, and the selection of class-balancing weights is important when dealing with imbalanced datasets to achieve a good performance.

7. Acknowledgments

This research was funded by an industrial doctorate grant of AGAUR between Universitat de Barcelona and Institut Cartogràfic I Geològic de Catalunya. This work was partially funded by projects RTI2018-095232-B-C21 (MINECO/FEDER, UE) and 2017SGR1742 (Generalitat de Catalunya).

References

- [1] Almeida, Danilo Roberti Alves de, et al. "The effectiveness of lidar remote sensing for monitoring forest cover attributes and landscape restoration." *Forest Ecology and Management* 438 (2019): 34-43.
- [2] Michałowska, Maja, and Jacek Rapiński. "A review of tree species classification based on airborne LiDAR data and applied classifiers." *Remote Sensing* 13.3 (2021): 353.
- [3] Armeni, Iro, et al. "3d semantic parsing of large-scale indoor spaces." *Proceedings of the IEEE CVPR conference*, 2016.
- [4] Chang, Angel X., et al. "Shapenet: An information-rich 3d model repository." 2015.
- [5] Zhou, Yin, et al. End-to-end multi-view fusion for 3d object detection in lidar point clouds. *Conference on Robot Learning. PMLR*, 2020.
- [6] Wu, Zhirong, et al. 3d shapenets: A deep representation for volumetric shapes. *Proceedings of the IEEE CVPR conference*, 2015.
- [7] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE CVPR conference*, 2017.
- [8] Qi, Charles Ruizhongtai, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in neural information processing systems* 30, 2017.
- [9] Yi, Li, et al. "Gspn: Generative shape proposal network for 3d instance segmentation in point cloud." *Proceedings of the IEEE CVPR conference*, 2019.
- [10] Landrieu, Loic, and Martin Simonovsky. "Large-scale point cloud semantic segmentation with super-point graphs." *Proceedings of the IEEE CVPR conference*, 2018.
- [11] Huang, Qiangui, Weiyue Wang, and Ulrich Neumann. "Recurrent slice networks for 3d segmentation of point clouds." *Proceedings of the IEEE CVPR conference*, 2018.
- [12] Zhao, Hengshuang, et al. "Point transformer." *Proceedings of the IEEE/CVF ICCV conference*, 2021.
- [13] Zoumpelas, T., et al. Benchmarking Deep Learning Models on Point Cloud Segmentation. *Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence*, 2021.
- [14] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." *Proceedings of the IEEE/CVF CVPR Conference*, 2019.
- [15] Cui, Yin, et al. "Class-balanced loss based on effective number of samples." *Proceedings of the IEEE/CVF CVPR conference*, 2019.
- [16] Van Horn, Grant, et al. "The devil is in the tails: Fine-grained classification in the wild.", 2017.
- [17] Pettorelli, N. *The normalized difference vegetation index*. Oxford University Press, 2013.
- [18] Terrasolid software. <https://geocue.com/software/terrasolid/>

Breast Tumor Classification in Digital Tomosynthesis Based on Deep Learning Radiomics

Loay HASSAN^{a,1}, Mohamed ABDEL-NASSER^{a,b}, Adel SALEH^c and Domenec PUIG^a

^a*Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

^b*Department of Electrical Engineering, Aswan University, Egypt*

^c*Gaist Solutions Ltd., Skipton BD23 2TZ, UK*

Abstract. Breast cancer is the most frequently diagnosed cancer in women globally. Early and accurate detection and classification of breast tumors are critical in improving treatment strategies and increasing the patient survival rate. Digital breast tomosynthesis (DBT) is an advanced form of mammography that aids better in the early detection and diagnosis of breast disease. This paper proposes a breast tumor classification method based on analyzing and evaluating the performance of various of the most innovative deep learning classification models in cooperation with a support vector machine (SVM) classifier for a DBT dataset. Specifically, we study the ability to use transfer learning from non-medical images to classify tumors in unseen DBT medical images. In addition, we utilize the fine-tuning technique to improve classification accuracy.

Keywords. Breast Cancer Classification, Digital breast tomosynthesis, Computer vision, Deep learning, Support Vector Machine

1. Introduction

Breast cancer is one of the most deadly deceases affecting females worldwide. The clinical studies have shown that early identification and classification of breast tumors has considerably improved the patient's treatments [1]. Generally, current common modalities for breast cancer screenings include mammography (X-ray images of the breast), breast ultrasound (BUS), thermograms, magnetic resonance imaging (MRI), and digital breast tomosynthesis (DBT) [2,3,4,5,6]. Digital breast tomosynthesis (DBT) is a promising new imaging modality for breast cancer screening that has the potential to overcome the limitations of traditional mammography. Instead of the projectional 2-dimensional images like in mammography, DBT delivers depth information through a practically 3-dimensional structural image of the breast volume (cross-sectional slices) and offers better performance [7]. DBT involves passing the X-ray tube in an arc over a fixed com-

¹Corresponding Author: Loay Hassan, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain; E-mail: loay.abdelrahimosa@urv.cat.

pressed breast at numerous angles to obtain a series of mammographic images [8]. Computer software subsequently reconstructs these individual mammographic images into a sequence of 3-dimensional and high-resolution slices. These 3-dimensional images notably decrease the influence of dense tissues, which can obscure or make it even harder to identify breast tumors [9].

While DBT can alleviate the limitation of tissue superimposition in mammography by providing superior tissue visibility, misdiagnosis is common, particularly in the early stages of breast cancer. So, a professional pathologist's experience and knowledge are required for a reliable breast cancer diagnosis. Regrettably, examining a more significant number of slices per breast volume raises clinical workflow issues accompanied by the fact that as the number of slices to examine grows, experts' monitoring of findings increases. Therefore, several computer-aided diagnostic (CAD) systems help clinicians identify breast tumors on various breast image modalities. In particular, with DBT images, the performance of the CAD systems will likely be better.

Recently, with the noteworthy development in the deep learning frameworks in many medical field studies [10], researchers have been inspired to investigate the application of deep learning to study the use of deep learning in developing efficient automated CAD systems for the breast cancer classification task [11]. For instance, Chougrad et al. [12] explored the performance of the three recent and powerful state-of-the-art CNN models to predict the correct diagnosis for various mammography breast cancer datasets. The obtained results demonstrate that the explored models are performant and can predict if the tumors are benign or malignant with accuracy ranging from 95 to 98% with different datasets. Adeyinka et al. [13] presented a discriminative fine-tuning deep learning-based method for breast cancer classification on mammograms. Their work aims to perform fine-tuning training using five popular pre-trained CNN models. Unlike traditional fine-tuning, which involves training the entire network again using a specified dataset, discriminative fine-tuning is introduced where they assign different learning rates and momentum to each layer of the network during the training process. The performance of their method is evaluated using the INBreast dataset [14] achieving the highest accuracy of 99% by the DensNet model.

For DBT images, Bevilacqua et al. [15] proposed a supervised deep learning-based normal/abnormal lesions classification method for breast tomosynthesis images. Practically, they compared the performance of two different classification approaches. The first approach is to utilize a shallow artificial neural network (ANN) classifier that takes morphological and textural hand-crafted features like the Grey Level Co-occurrence Matrix (GLCM) as input. The second approach is based on automatically computed features and extracting several sets of features using deeper CNN from DBT images. The final results showed that the second classification approach performs better. With a private DBT dataset, they obtained an accuracy of 92% with the VGG network. Samala et al. [16] developed a multi-stage deep learning framework for classifying malignant and benign tumors in DBT images. The main idea of that work is to study the effectiveness of the transfer learning approach when a pre-trained CNN model on non-medical images is first fine-tuned to a related task in the medical imaging domain before being fine-tuned to the target task in an attempt to overcome the lack of large training data. In practice, the proposed framework consists of two stages. In the first stage, they utilized the AlexNet model trained by more than a million non-medical images from the ImageNet dataset to be fine-tuned with less than 3000 patch images extracted from mammograms. Then,

they used the fine-tuned model from the first stage to further fine-tuned with less than 1500 patches from DBT images in the second stage. The experimental results evaluated on a private DBT dataset showed that the accuracy of multi-stage transfer learning is improved by 6% over single-stage transfer learning.

Zhang et al. [17] presented breast normal/abnormal lesions classification method in DBT images where a traditional 2-D deep CNN model is operated on the whole volume of 3-D DBT images, regardless of the number of slices. The fundamental idea behind their work is that instead of only using small tumor patches, they focus on full-image classification. Specifically, for z-slices of the DBT image, every three consecutive slices are stacked as a three-channel image input to the feature extractor network. Then, they generated a feature map by pooling the features extracted for the binary classification. The experimental results evaluated using a private clinical DBT dataset showed that classification performed best using the AlexNet model as feature extractor with MaxPooling for feature fusion.

Although various breast tumor classification methods in DBT are proposed, most of them are limited to classifying normal and abnormal lesions. Also, most of these proposed studies are evaluated using a private DBT dataset. Therefore, automated benign/malignant tumor classification of the breast in DBT still faces several challenges due to the lack of public available DBT datasets. In this paper, based entirely on the only publicly available DBT dataset, we present benign/malignant tumor classification using a support vector machine (SVM) classifier and deep learning classifier for DBT images based on radiomics extracted from well-known CNN classification models. We explore the ability to use transfer learning from non-medical images to classify tumors in DBT images. Alongside, the fine-tuning approach is applied to improve the classifier's efficiency.

The remainder of this paper is designed as follows. Section 2 describes the proposed classification method. The experimental results and discussion are presented in section 3. Section 4 concludes the paper.

2. Methodology

Figure 1 shows the proposed breast tumor classification method. The key elements of the proposed method are data preparation, deep learning-based feature (Radiomics) extraction, and classification heads. Below, we describe the proposed method in detail.

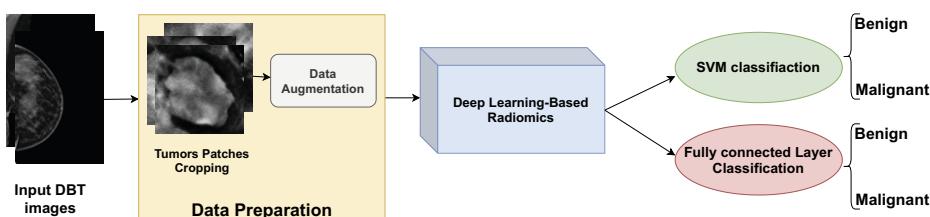


Figure 1. Breast tumor classification in DBT images based on deep learning-based radiomics.

2.1. Data Preparation

In this study, the data preparation step consists of two stages: the tumor patch cropping stage and the data augmentation stage. It should be noted that we used the only publicly available DBT images dataset named the DBTex challenge dataset [18] which contains 1000 breast tomosynthesis scans from 985 patients. Disastrously, not all images are fully annotated, as out of the 101 patients, only 224 DBT images have annotations leaving us with few usable DBT images.

For the tumor patch cropping stage, in terms of ROIs selection, tumors are center cropped and extracted from the annotated DBT image. The selected ROIs are then all resized into identical dimensions of 224×224 , to fit the input of deep learning networks in the feature extraction step. Then, we divide the DBT images (patient-wise) and the corresponding patches into training and testing sets as shown in Table 1.

Regarding data balance, we seek to achieve that the training and the test set involve the

Table 1. Overview of the DBT dataset.

	No. of patients		No. of tumor patches	
	Train	Test	Train	Test
Benign	50	12	120	23
Malignant	27	12	63	23
Total	77	24	183	46
Augmented Data	77	24	246	46

equivalent amount of two classes of tumors. Therefore, in the data augmentation stage, the number of training data tumor patches increases (see Table 1). In particular, to balance the benign and malignant number of patches in the training set, we doubled the number by jointly flipping all malignant patches in the training set horizontally and vertically. This eventually results in 120 benign tumor patches and 126 malignant tumor patches, with total patches of 246 in the training set, in addition to 23 benign tumor patches and 23 malignant tumor patch, with a total of 46 in the test set, which achieves a ratio of approximately 80% of the training images and 20% of the test images of the whole set of images.

2.2. Deep learning-based radiomics

In the breast tumor classification task, each tumor input patch supposes to be classified as benign or malignant. However, in the case of automated machine-based classifiers, this task is much more difficult due to the need to characterize those input images into discriminative features. In contrast to the traditional methods based on hand-crafted feature extraction, deep learning models enable robust and automated feature extraction.

As shown in Figure 2, the radiomics extraction process has been carried out in two phases. In the first phase, a transfer learning approach, classification CNN models pre-trained on non-medical images are used to directly extract features from the medical DBT images to train one of the classifiers in the classification heads. Since these models have been trained on a dataset with many images, for example, trained on the ImageNet dataset, which contains 1.2 million non-medical images for a 1000 class image classification problem, they can extract meaningful features that can be used for direct classifi-

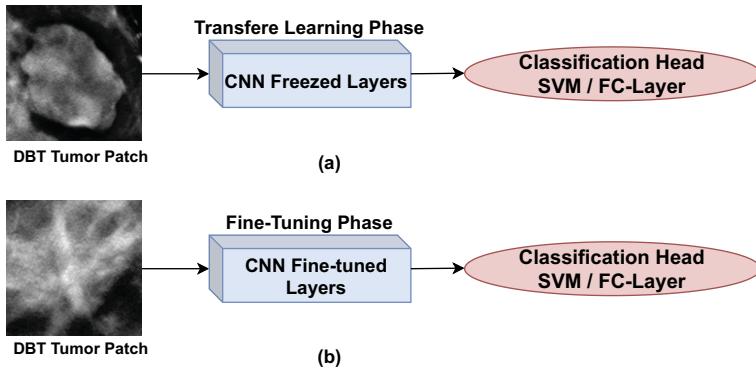


Figure 2. Feature extraction approaches: (a) transfere learning, and (b) fine-tuning.

cation.

In the second phase, a fine-tuning approach, classification CNN models pre-trained on non-medical images are fine-tuned on the training set of 246 DBT images. Applying fine-tuning allows us to utilize the pre-trained networks to the related breast tumor classification task in the DBT medical imaging domain, leading to higher accuracy. For both feature extraction phases, we investigate the performance of various of the most innovative deep learning models inspired by the well-known CNN backbone architectures such as AlexNet [19], VGG [20], ResNet [21], WideResNet [22], SqueezeNet [23] and EfficientNet [24] to extract discriminative features from DBT images.

2.3. Classification heads

Lastly, the generated tumor radiomics from the feature extraction stage can be classified by various machine learning classification algorithms. In this paper, we employ the classification process through two scenarios: end-to-end deep learning classification by a fully connected (FC) layer and an SVM classifier.

In the case of the end-to-end deep learning classification, we modify the classification part of the last FC-layer to the set of classes of our task, i.e., two classes of benign and malignant. Therefore, for the transfer learning approach, we do not need to re-train the entire model, just the final classification part, which will be trained from scratch on top of the pre-trained model. Differently, the fine-tuning approach jointly trains both the newly-modified classifier layers and the whole layers of the base model.

In this study, we also employ the SVM algorithm, a fast and dependable classification algorithm that performs very well with a limited amount of data to analyze, making it suitable for the DBT breast tumor classification task due to the lack of data. Accordingly, whether the radiomics are extracted using transfer learning or fine-tuning approaches, these deep features are used to train the SVM classifier.

2.4. Implementation

To train the end-to-end deep learning classifier for both transfer learning and fine-tuning approaches, we use Adam to optimize the evaluated CNN model with a learning rate of $1e - 4$. It should be noted that all models were trained for 30 epochs. On the other hand,

to train the SVM classifier, radiomics are extracted first using one of the approaches mentioned above (i.e., transfer learning or fine-tuning), then extracted features are used as input for training and evaluating the SVM classifier. For fair comparisons, the proposed evaluated classifiers were trained using the same 246 DBT images and tested with the same 46 DBT images. All the experiments were performed using the Pytorch framework using a 64-bit Ubuntu operating system with 3.6 GHz intel core i7 with 32GB of RAM and Nvidia RTX3080 with 10GB of video RAM.

3. Experimental Results and Analysis

3.1. Performance evaluation of transfer learning approach

Table 2 presents a quantitative comparison between the SVM classifier and the Deep learning classifier trained with radiomics extracted from the evaluated models (AlexNet, VGG, ResNet, wideResNet, SqueezeNet, and EfficientNet) for the transfer learning approach in terms of percentage accuracy.

As one can see, the SVM classifier trained with radiomics extracted from the pre-trained

Table 2. The performance of the SVM classifier and deep learning classifier with transfer learning approach.

Model	SVM classifier	Deep learning classifier
AlexNet	63.04%	56.52%
VGG19	65.22%	54.35%
ResNet50	52.17%	60.87%
WideResNet101	60.87%	60.87%
SqueezeNet	56.52%	63.04 %
EfficientNet	63.04%	58.70%

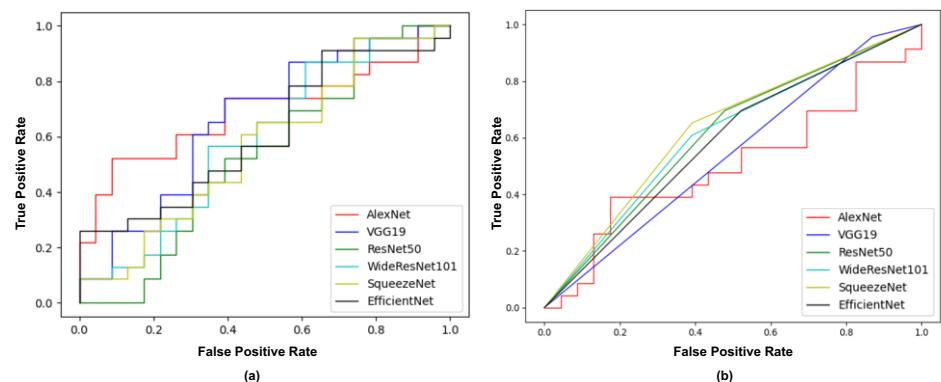


Figure 3. ROC curves of (a) the SVM, and (b) DL classifiers.

VGG model achieved the best classification results with obtained accuracy of 65.22% compared to the end-to-end deep learning classifier. Meanwhile, the deep learning classifier trained with radiomics from pre-trained SqueezeNet obtained a competitive classification accuracy of 60.04%. From the overall results in Table 2, we can say that for the

transfer learning approach where deep learning models trained on non-medical data and are used to extract features from unseen DBT images, the SVM classifier has a promising classification accuracy and surpassed Deep learning classification on approximately all evaluated model.

Figure 3 shows the ROC curves for SVM and deep learning classifier for the transfer learning approach of all the evaluated pre-trained models. It is shown that the SVM classifier trained with VGG deep features (radiomics) performs best with an AUC of 0.65 compared to other feature extraction models. The second-best performing classifier is the deep learning classifier trained with wideResNet radiomics, even if its overall accuracy is lower than features from SqueezeNet with an AUC of 0.67.

Regarding class classification accuracy, Figure 4 presents the confusion matrix for the

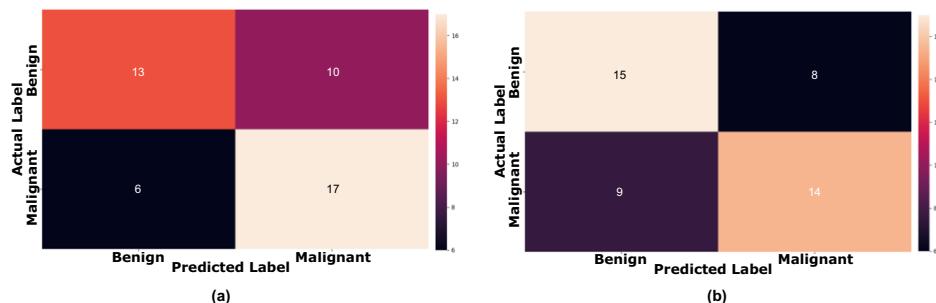


Figure 4. Confusion matrices of (a) the SVM classifier trained with deep features of VGG19 Network and (b) the DL classifier trained with deep features of SqueezeNet Network.

best SVM classifier with radiomics from the pre-trained VGG model and the best end-to-end deep learning classification with radiomics from the pre-trained SqueezeNet model. As one can see, the SVM classifier succeeded in classifying malignant tumors by more than 73%, outperforming the deep learning classifier by approximately 13%, while the deep learning classifier outperformed the SVM classifier in classifying benign tumors by 4%.

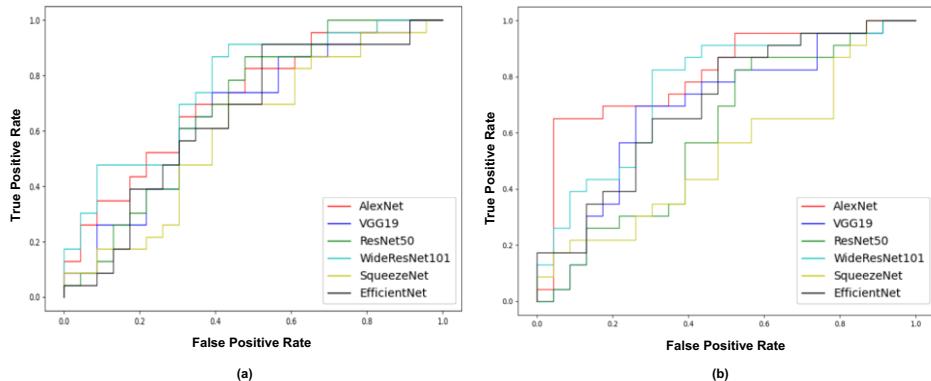
3.2. Performance evaluation of fine-tuning approach

In the case of the fine-tuning method, Table 3 compares the classification performance of the SVM classifier and the end-to-end deep learning classifier for the evaluated fine-tuned deep feature (Radiomics) extractor models trained with the training set of the DBT images dataset. As shown, when comparing values from Table 2 to values from Table 3, we can argue that fine-tuning the deep learning-based radiomics can yield noticeable improvements in terms of classification accuracy for both classifiers, especially for the end-to-end deep learning classifier.

With the fine-tuning approach, the performance of the SVM classifier trained with radiomics from the AlexNet model increased by 8.7%. Besides, the classification accuracy of the end-to-end deep learning classifier of the AlexNet model was also advanced by 23.91%. It is also noticeable that there is a comparative performance in the results of the deep learning classifier for the rest of the models, with classification accuracy ranging from 65.22% to 73.91%.

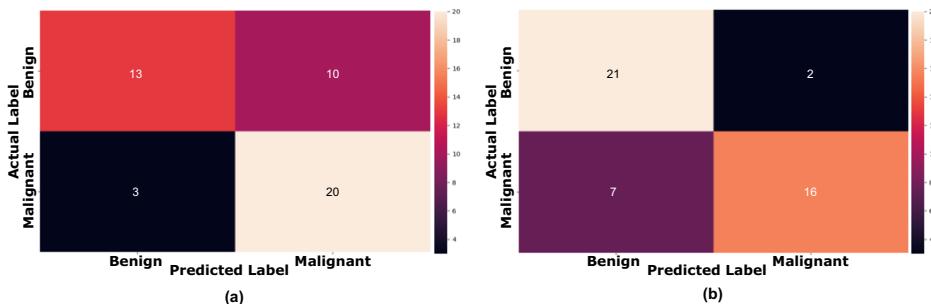
Table 3. The performance of the SVM classifier and deep learning classifier with the fine-tuning approach.

Model	SVM classifier	Deep learning classifier
AlexNet	71.74 %	80.43%
VGG19	60.87%	67.39%
ResNet50	52.17%	65.22%
WideResNet101	50.00%	73.91%
SqueezeNet	60.87%	69.57%
EfficientNet	60.87%	67.39%

**Figure 5.** ROC curves of (a) the SVM, and (b) DL classifiers.

The ROC curves for SVM and deep learning classifier for the fine-tuning method are shown in Figure 5. It is shown that the SVM classifier and the deep learning classifier trained with radiomics from the AlexNet model have the best performance, confirming the high efficiency of both classifiers presented in Table 3 obtaining AUC of 0.72 and 0.80, respectively.

Regarding class classification accuracy, Figure 6 shows the confusion matrix for the best

**Figure 6.** Confusion matrices of (a) the SVM classifier trained with deep features of AlexNet Network and (b) the DL classifier trained with deep features of AlexNet Network.

SVM classifier with radiomics from the fine-tuned AlexNet model and the best end-to-end deep learning classification with the fine-tuned AlexNet model. Here, the deep learning classifier has better accuracy in classifying the benign tumors with obtained accu-

racy of 91%, while the SVM classifier outperformed in classifying Malignant tumors by 17%. Based on the above analysis, we can conclude that to transfer learning techniques to extract features in cooperation with the SVM classifier is better for DBT images than the deep learning classifier. In contrast, end-to-end deep learning classification based on radiomics from fine-tuned models can significantly improve breast tumor classification accuracy. Of note, the classification results could be further improved by utilizing an optimization algorithm, such as the stochastic whale optimization algorithm [25] to find the optimal parameters of SVM.

4. Conclusions

This paper presents a breast tumor classification method for digital breast tomosynthesis images (DBT) based on radiomics extracted from the most innovative deep learning classification models. Our work first investigated the transfer learning technique where pre-trained on non-medical images CNN models are used to directly extract radiomics from the medical DBT images to train support vector machine and deep learning-based classifier. Secondly, the fine-tuning technique where the CNN models pre-trained on non-medical images is fine-tuned on the training set of DBT images. With the only publicly available digital breast tomosynthesis dataset, our experiments showed that end-to-end deep learning classification with radiomics extracted from the fine-tuned AlexNet model achieved the best classification accuracy of 80.43%.

The future work will focus on developing a breast tumor classification method based on the aggregation of robust deep learning-based feature extraction models.

Acknowledgement

The Spanish Government partly supported this research through Project PID2019-105789RB-I00.

References

- [1] Gunjan Chugh, Shailender Kumar, and Nanhay Singh. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation*, 13(6):1451–1470, jan 2021.
- [2] Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, and Domenec Puig. The impact of pixel resolution, integration scale, preprocessing, and feature normalization on texture analysis for mass classification in mammograms. *International Journal of Optics*, 2016.
- [3] Mohamed Abdel-Nasser, Antonio Moreno, and Domenec Puig. Temporal mammogram image registration using optimized curvilinear coordinates. *Computer methods and programs in biomedicine*, 127:1–14, 2016.
- [4] Vivek Kumar Singh, Mohamed Abdel-Nasser, Farhan Akram, Hatem A Rashwan, Md Mostafa Kamal Sarker, Nidhi Pandey, Santiago Romani, and Domènec Puig. Breast tumor segmentation in ultrasound images using contextual-information-aware deep adversarial learning framework. *Expert Systems with Applications*, 162:113870, 2020.
- [5] Mohamed Abdel-Nasser, Adel Saleh, Antonio Moreno, and Domenec Puig. Automatic nipple detection in breast thermograms. *Expert Systems with Applications*, 64:365–374, 2016.
- [6] JAMES V. FIORICA. Breast cancer screening, mammography, and other modalities. *Clinical Obstetrics and Gynecology*, 59(4):688–709, dec 2016.

- [7] Srinivasan Vedantham, Andrew Karella, Gopal R. Vijayaraghavan, and Daniel B. Kopans. Digital breast tomosynthesis: State of the art. *Radiology*, 277(3):663–684, dec 2015.
- [8] Xuan-Anh Phi, Alberto Tagliafico, Nehmat Houssami, Marcel J. W. Greuter, and Geertruida H. de Bock. Digital breast tomosynthesis for breast cancer screening and diagnosis in women with dense breasts – a systematic review and meta-analysis. *BMC Cancer*, 18(1), April 2018.
- [9] Kathryn P. Lowry, Rebecca Yates Coley, Diana L. Miglioretti, Karla Kerlikowske, Louise M. Henderson, Tracy Onega, Brian L. Sprague, Janie M. Lee, Sally Herschorn, Anna N. A. Tosteson, Garth Rauscher, and Christoph I. Lee. Screening performance of digital breast tomosynthesis vs digital mammography in community practice by patient age, screening round, and breast density. *JAMA Network Open*, 3(7):e2011792, July 2020.
- [10] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3):47, August 2018.
- [11] Parita Oza, Paawan Sharma, Samir Patel, and Pankaj Kumar. Deep convolutional neural networks for computer-aided breast cancer diagnostic: a survey. *Neural Computing and Applications*, 34(3):1815–1836, jan 2022.
- [12] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer Methods and Programs in Biomedicine*, 157:19–30, apr 2018.
- [13] Adeyinka P. Adedigba, Steve A. Adeshina, and Abiodun M. Aibinu. Performance evaluation of deep learning models on mammogram classification using small dataset. *Bioengineering*, 9(4):161, apr 2022.
- [14] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. INbreast. *Academic Radiology*, 19(2):236–248, feb 2012.
- [15] Vitoantonio Bevilacqua, Antonio Brunetti, Andrea Guerrero, Gianpaolo Francesco Trotta, Michele Telegrafo, and Marco Moschetta. A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cognitive Systems Research*, 53:3–19, jan 2019.
- [16] Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Caleb D. Richter, and Kenny H. Cha. Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, 38(3):686–696, mar 2019.
- [17] Yu Zhang, Xiaoqin Wang, Hunter Blanton, Gongbo Liang, Xin Xing, and Nathan Jacobs. 2d convolutional neural networks for 3d digital breast tomosynthesis classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, nov 2019.
- [18] SPIE-AAPM-NCI DAIR Digital Breast Tomosynthesis Lesion Detection Challenge. <https://www.aapm.org/GrandChallenge/DBTex/>. Accessed: 2022-05-20.
- [19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.
- [20] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, November 2015.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.
- [22] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016.
- [23] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 0.5mb model size, 2016.
- [24] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.
- [25] FatmaAlzahra Mohamed, Mohamed Abdel-Nasser, Karar Mahmoud, and Salah Kamel. Economic dispatch using stochastic whale optimization algorithm. In *2018 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 19–24. IEEE, 2018.

Automatic Outdoor Image Geolocation with Focal Modulation Networks

Fabio MURGESE^{a,1}, Gerard ALCAINA^{a,2}, Mehmet Oğuz MÜLÂYİM^a,
Jesus CERQUIDES^a, and Jose Luis FERNANDEZ-MARQUEZ^b

^aArtificial Intelligence Research Institute (IIIA), CSIC, Cerdanyola del Vallès, Spain

^bCitizen Cyberlab, University of Geneva, Geneva, Switzerland

Abstract.

We address the problem of estimating a photo's geographical location. Success in this estimation enables many impactful applications, like facilitating Disaster Management circumstances. However, this is also a very challenging task. Due to the complexity of the problem, we restrict the area of geolocation to a single city, treating geolocation as a classification problem where the districts of a city are the classes to be distinguished. In this paper, we exploit the *Focal Modulation Network* that is proven to perform effectively and efficiently in visual modeling for real-world applications. Experimental results on two diverse datasets, crawled from online sources, show the effectiveness of our approach. We can geolocate correctly more than two-thirds of test images from the larger dataset and about one-third from an experimental training dataset of a ten-times smaller size.

Keywords. Image Classification; Focal Modulation Networks; Outdoor Image Geolocation; Barcelona Geolocation

1. Introduction

Image geolocation (i.e., identifying the geographical location of an image) is a highly challenging task since photos taken even from the same location exhibit immense variations due to different camera settings, seasons, daylight conditions, and present objects. Also, images are often ambiguous and could provide very few cues about their location. In the absence of discriminative landmarks, humans can leverage their world knowledge to infer the location of a photo, using hints like the language of street signs or the driving direction of cars. Most previous work on image geolocation focused on identifying and geolocating landmark buildings (e.g., [3, 14]) whereas very few approaches tried to geolocate images just by using pixels (e.g., [7, 16]).

Our primary motivation for precise outdoor image geolocation is its application to the Disaster Management field, where it can have a critical impact because a fast response is of paramount importance to help emergency aid. Social media data has demonstrated to be extremely relevant to evaluate damage and improve the understanding after a natural disaster occurs [4, 5], especially, in the first 24/48h which are crucial to allow emergency

¹Corresponding Author: Fabio Murgese, IIIA, CSIC, Cerdanyola 08193. E-mail:fabiomur95@gmail.com

²Corresponding Author: Gerard Alcaina, IIIA, CSIC, Cerdanyola 08193. E-mail:galcaina98@gmail.com

responders to coordinate their actions. Finding the location of the social media content is a major challenge to make social media information ready to be used by emergency responders. For example, it is relevant to observe a Twitter photo containing a school which has been damaged after an earthquake, however that information will be hardly usable if we do not know the location where the photo was taken, i.e., where the damaged school is located.

Online image databases and social media provide us with invaluable data for training our Machine Learning models and later quickly geolocating the images shared from disaster areas. In this work, we propose to exploit the *Focal Modulation Network* (FocalNet) [19] which is demonstrated by its authors to outperform the state-of-the-art for effective and efficient visual modeling in real-world applications. The main contribution of this paper is an open-source image crawler software that is able to retrieve all available data from Flickr³ and Mapillary⁴ for a chosen city and attach their district information in order to fit with FocalNets requirements for the training of the geolocator. Consequently, we can discriminate between the areas of a city at different granularities, a non-trivial task within the geolocation field. Specifically, we can choose to distinguish districts or neighborhoods by using alternative data sources, in our case, stored as Geo-JSON⁵ files that contain the coordinates of the geographical borders of cities and their sub-regions.

Additionally, we publish two datasets that comprise the coordinates and the districts of images that were taken in Barcelona, Spain: 1) The Flickr dataset is made of about 18k images that were crawled by our pipeline posing the districts of the city of Barcelona as query keywords; 2) The Mapillary dataset consists of more than 182k images, with attached geographic coordinates and district information. The photos in the latter dataset were crawled by a recursive algorithm that enabled us to query the Mapillary Application Programming Interface (API) iteratively without worrying about the limit of data that can be retrieved in one shot from the platform. More details will be disclosed in Section 3. Then, we trained different FocalNet models on both datasets to understand which benefits come with the usage of two datasets of very different nature: one large robotic and one smaller, non-robotic.

We introduce related work in Section 2. In Section 3, we preface the process followed to gather data from multiple sources. Then, in Section 4, we give the set-up for experiments and reflect on the performances of trained models and the results. Section 5 summarizes the findings of this work and gives the outlook for the next research steps.

2. Background

In the field of emergency management, social media images have been geolocated by humanitarian communities such as GISCorp⁶, VOST Europe⁷ and Standby Task Force (SBTF)⁸. Humanitarian networks involved hundred of volunteers contributing remotely.

³<https://www.flickr.com/>

⁴<https://www.mapillary.com/>

⁵<https://geojson.org/>

⁶<https://www.giscorps.org/>

⁷<https://vosteurope.org/>

⁸<https://standbytaskforce.org/>

Crowdsourcing participatory platforms have been designed to simplify the manual geolocation of social media images [15]. However, automatic geolocation would help the crowd effort scale, ease the work of the communities by allowing them to focus on only the refinement of the automatically calculated geolocation.

In recent years, there are two main approaches that tackle the problem of geolocating images: by comparison and by classification [12]. The former is based on an information-retrieval approach that matches a query photo against millions of geotagged images, whereas the latter tries to predict the right class using a single trained model.

Initially, Im2GPS [6] attempted to solve the problem by retrieving the neighbors of a query photo in a database of 6 million geotagged Flickr images and geolocating the query by assigning it the location of the nearest match. Nevertheless, due to the difficulty to classify non-generic scenes using this technique, multiple alternatives appeared addressing the problem with image classification.

Later, PlaNet [18] approached the problem in a different way: it is a worldwide image geolocation classifier that divides the Earth surface into multi-scale geographic cells. The main drawback of this technique was the difficulty to cover places where photos are very unlikely to be taken. New techniques inspired by this approach, such as [9, 13], were proven to get better contextual information of the images adding a hierarchical model and scene classifiers.

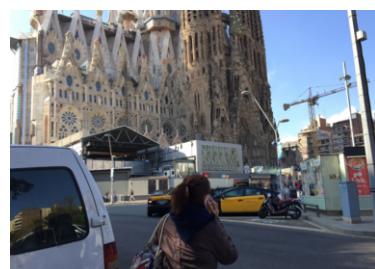
The novelty introduced by this paper is to use FocalNets [19] within the geolocation context, adopting a classification approach, as they have been proven to outperform the state-of-the-art Self-Attention counterparts [8, 17]. Concisely, focal modulation first aggregates contexts around each query, in order to be able to modulate the query with the combined context. In this way, it simplifies the process by enabling input-dependent token (i.e., the query) interaction. Also, with this model it is possible to generate summarized tokens at distinct levels of granularity applying query-agnostic aggregations. In the end, these contexts are fused into the query vector, after being selectively aggregated in conformity with the query content.

3. Data gathering

When dealing with the outdoor image geolocation problem, first step is to understand the nature of the data that can help solve the task at hand. More specifically, the sources of datasets for geolocation can be categorized as robotic or non-robotic. Robotic ones are



(a) Extracted from Flickr.

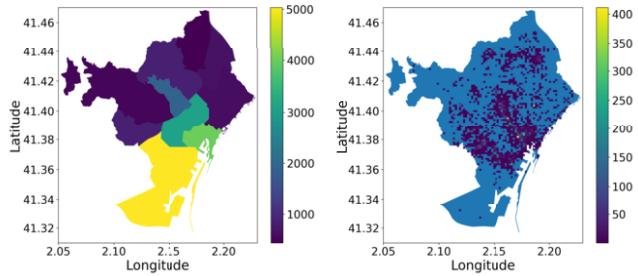


(b) Extracted from Mapillary.

Figure 1. Sample images of non-robotic (left) and robotic (right) styles.

District	Images
Sants-Montjuïc	5,070
Ciutat Vella	4,017
Eixample	3,170
Gràcia	1,888
Les Corts	918
Horta-Guinardó	864
Sant Andreu	651
Sant Martí	603
Sarrià-Sant Gervasi	525
Nou Barris	463

(a) # of images.



(b) Distribution of images.

(c) Exact image locations.

Figure 2. Barcelona Flickr crawled dataset.

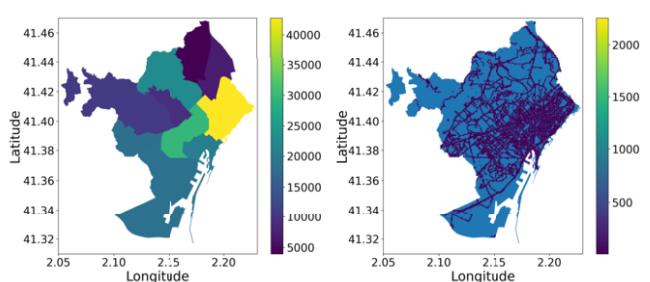
typically created by recording videos from static cameras mounted on a vehicle (e.g., a car, a bicycle), resulting in images available in sequences temporally consistent (assuming the motion of the vehicle) and from a steady point of view (usually the street), with no big changes in the vertical direction, as explained in [12]. Non-robotic datasets, on the other hand, are created by collections of online images, usually taken nearby points of interest, making the geolocation task easy for landmarks, but difficult for not widely-known places. And, here comes the potential value of exploiting robotic datasets, which, by nature, have better coverage of these less considered areas. Figure 1 gives a flavour of the diversity of both styles.

In our case, we use Mapillary to assemble robotic data and Flickr for non-robotic data. Public APIs of these online imagery platforms help us retrieve the maximum amount of images of rich variety for a specific region. Once we retrieve data for a specific city from the two sources, we attach the district information for every image leveraging a GeoJSON file for the city of interest, thus facilitating the replicability of the pipeline for other cities.

The first dataset we curated is non-robotic and consists of images crawled from Flickr. To query the Flickr API, we used the words representing the ten districts of Barcelona and the word “barcelona” itself. Due to its API limit of maximum 4,000 images per query word, the resulting Flickr dataset is much smaller than the Mapillary one,

District	Images
Sant Martí	42,735
Eixample	29,720
Horta-Guinardó	22,520
Ciutat Vella	20,298
Sants-Montjuïc	18,345
Les Corts	18,053
Sarrià-Sant Gervasi	10,545
Gràcia	9,106
Sant Andreu	6,864
Nou Barris	3,944

(a) # of images.



(b) Distribution of images.

(c) Exact image locations.

Figure 3. Barcelona Mapillary crawled dataset.

and contains 17,939 images. Figure 2 gives the distribution of Flickr images in our area of interest.

Then, we created our second dataset by crawling Mapillary data posing as queries coordinate boxes within the geographical borders of the city of Barcelona. Specifically, we used as a starting point the centroid of the city, computed using the GeoJSON coordinates, and crawled the photos available within a box around the centroid. Having a limitation of 2,000 images imposed by the Mapillary API, we needed to create an algorithm that splits our original box recursively into multiple smaller ones to manage to retrieve all existing images within the municipality. In the end, we created a robotic dataset of 182,130 images, representing the city of Barcelona. The distribution of the images in this dataset are given in Figure 3.

From Figures 2 and 3, we can clearly see the differences between two datasets regarding both the distribution of images, and, most importantly, their densities. Figure 3b displays more uniformly-distributed images throughout the surface of the whole municipality of Barcelona, with a peak in Sant Martí due to a denser road infrastructure that can be seen in Figure 3c, where the image spots highlight the city roads. This contrasts with Figure 2b where we have a higher number of images concentrated into what corresponds to the most well-known places of the city as it's shown on the hot spots of Figure 2c. Both datasets and the code for generating them are publicly available at [2] and [1], respectively.

4. Experiments

Within the scope of this paper, we first ran our experiments on the smaller Flickr dataset, then we compared the results with the larger robotic Mapillary dataset, using FocalNets to discriminate the different districts of Barcelona. First, using the city's GeoJSON data, we assigned each image in the dataset to the class representing the district in which the image's GPS coordinates fall. Then, we split both datasets into training and validation sets of 90% and 10% sizes respectively.

Yang et al. [19] made available three different models: FocalNet-T (tiny), FocalNet-S (small) and FocalNet-B (base). These models differ in depth layouts and hidden dimensions. For our first experiment, we used the FocalNet-S model and we trained this network from scratch on the 17,939 Flickr images that were previously cropped to 224×224 pixels. We trained⁹ the model for 50 epochs with a batch size of 32 using the default hyperparameters configuration. This includes the AdamW optimization [10] and a cosine learning rate scheduler [11], with initial learning rate $5e-7$ for the first 20 warm-up epochs and $5e-4$ for the following ones. Gradient clipping norm is set to 5.0 and the weight decay is set to 0.05.

We carried out a second experiment with the same dataset exploiting the FocalNet-T model using the same configuration as the previous one in order to compare how the two models behave with the geolocation task at hand.¹⁰ The code for the experiments is publicly available at [1].

⁹Experiments are run on an AMD EPYC 7313P 16-Core processor with 128GB RAM and an NVIDIA GeForce RTX 3090 graphic card.

¹⁰We note that we left experiments with the larger FocalNet-B model as future work due to time constraints and the limited availability of the shared resources to run the experiments.

Table 1 summarizes the classification accuracy we achieved with both models. We evaluated the models on the validation set and we reached a mean accuracy of 32.8% with the FocalNet-S model, while we achieved a slightly lower value with the FocalNet-T model. If we consider, instead, the accuracy of top-5 predictions, the score goes up to about 83% for both models, a much higher score that shows the uncertainty of the model predicting exactly the best class. We also observe that the training time drops considerably for the FocalNet-T model due to the model layout and less number of parameters to learn.

Considering our Flickr dataset, having a standard guess in favor of the most represented class (Sants-Montjuïc) we would have an accuracy of $\sim 28\%$. Now, comparing this score with the mean accuracy of our model (over all classes), we can argue that the model is learning, even by using this small and unbalanced dataset. These preliminary results with the small dataset encouraged us for the next experiments with the larger Mapillary dataset for Barcelona.

We carried out two different experiments using Mapillary data training both the FocalNet-T and FocalNet-S models for 50 epochs, using a batch size of 32 and the same hyperparameters as before. And as expected, we reached a much higher accuracy score using the larger dataset, as depicted in Table 1. As we saw with the smaller Flickr dataset, the smaller FocalNet-T model is as accurate as the FocalNet-S model for this task, but having the advantage of considerably smaller training time. We also give the validation cross-entropy loss and the accuracy of the top-1 predictions of all trained models in our experiments in Figure 4.

	Dataset	Accuracy top-1	Accuracy top-5	Training time
FocalNet-T	Flickr	32.1%	83.5%	38mins
FocalNet-S	Flickr	32.8%	83.1%	1h 33mins
FocalNet-T	Mapillary	68.2%	94.6%	9h 08mins
FocalNet-S	Mapillary	68.5%	94.6%	15h 27mins

Table 1. Barcelona district classification accuracies with FocalNets.

One thing to notice is that the Mapillary dataset results are much better than the Flickr ones. We argue that this could not be just because of the different sizes of the two datasets. We think that the much higher scores in the Mapillary based experiments could be due to the nature of the data itself: robotic datasets are the result of sequences

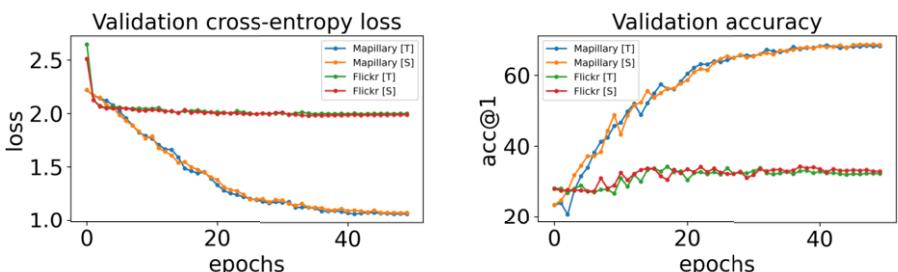
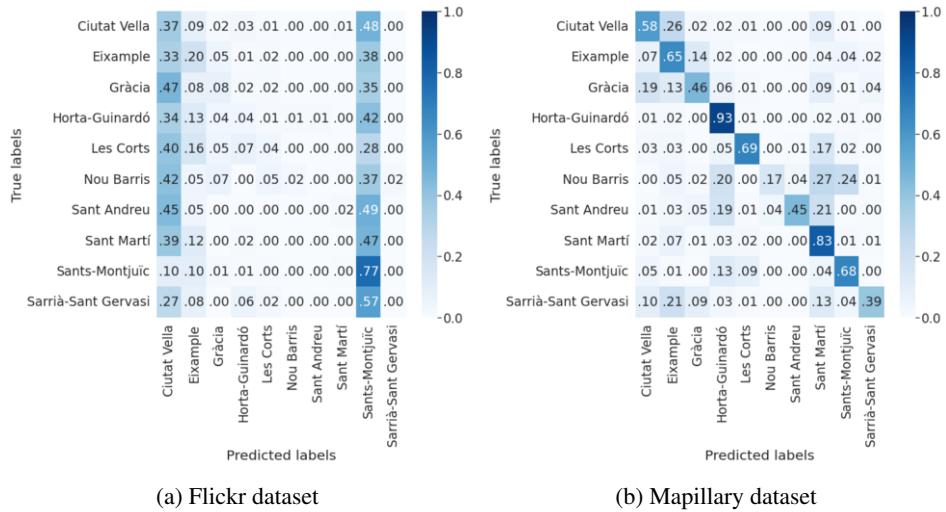


Figure 4. Validation curves on both datasets with FocalNet-T and FocalNet-S models.

of images from a consistent point of view, taken at different time spans. Therefore, some of the validation images can be relatively similar to the training ones: this dependency is inherited by how the photos have been shot in the first place. So, it could be possible that our models were trained on some images that are the previous sequences of frames used in our validation set to evaluate the models. In spite of everything, as we can see in Figure 4, the Mapillary validation losses and accuracies are likely to improve: with a higher number of epochs we could reach better scores.

Additionally, we generated confusion matrices over the validation data in order to compare the strengths and weaknesses of the geolocator obtained on both datasets. From Figure 5a, we can conclude that the geolocator for Flickr is prone to classify the images in a skewed distribution in favor of the two most represented classes. This contrasts with Figure 5b, where we observe that the classification process with the Mapillary dataset is quite robust with the exception of a minority of the classes.



(a) Flickr dataset

(b) Mapillary dataset

Figure 5. Confusion matrices on validation data with FocalNet-T models.

Our experiments show that, from a practical point of view, the classification using the Flickr dataset is not reliable enough; this could be because of the nature of the images that make up the dataset or as a result of the small dataset size and unbalancedness of the classes. We regard these first results as a motivation for further experimentation with non-robotic datasets.

5. Conclusions and future work

The motivation for our research regarding this paper is the automatic geolocation of images taken from disaster areas. In Disaster Management, timely spotting of the places affected by a calamity has a significant role in redirecting emergency help. To this end, correctly geolocating images taken from the affected area using an automated software pipeline could be of great aid to first responders and could facilitate the rescue of people subject to natural disasters. The replicability of such a pipeline in new disaster areas

could allow these emergencies to be managed with the least effort. For this purpose, by developing an automatic geolocator for a given area of interest, (e.g., a city, a region) we are providing a tool that could add impactful help even into dramatic situations.

As a first step to building a geolocation pipeline, we implemented a crawler for two major online image sources for cities, namely Flickr and Mapillary. As the second step, we chose a novel classification model, Focal Modulation Networks (FocalNets) [19], that outperforms the state-of-the-art and requires relatively short training time. We used FocalNets within the context, first, of a small non-robotic dataset and, second, of a larger robotic dataset.

We regard the uncertainty in predicting the best class, especially in the case of non-robotic images, as a symptom of the need for further investigation. The first thing we have in mind is to experiment with equally-sized non-robotic and robotic datasets to scrutinize the discrepancy between the first results with these two datasets. Second step will be to incorporate a k-fold cross validation in our pipeline. This stage would give us a better insight into the generalization capabilities of our models. A later step will be to use both our datasets in a hybrid fashion to train our geolocators in order to see their classification performance with mixed image types. Having success with such a hybrid dataset will give great value to the abundant robotic imagery available online (~ 1.5 billion street-level images available from all around the world within Mapillary) and would bring about great opportunities to leverage everyday-growing pieces of information that social media provide us with. This would mean that, with our pipeline, scraping the Mapillary and Flickr APIs for a new city and training the model with this new data will provide us with the opportunity to discriminate between the districts of a city of interest in relatively short amount of time. Moreover, the usage of data gathered from other social media platforms (e.g., Twitter) could be very insightful to test how the geolocator responds to disparate variety of images extracted from real-world emergencies. Thinking about Disaster Management, when people constantly post new images showing a different perspective of the after-effects, nearly real-time geolocation could be of great aid, even for cities never analyzed before.

The code for our pipeline and the experiments, and the datasets are publicly available at [1] and [2], respectively.

Acknowledgements This work was partially funded by the EU H2020 project Crowd4SDG “Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience”, #872944. Fabio Murgese is an MSc student of the Computer Science department at the Università di Pisa and Gerard Alcaina is an MSc student of the Mathematics department at the Universitat Autònoma de Barcelona.

References

- [1] Artificial Intelligence Research Institute (IIIA), CSIC. Code for image geolocation. <https://github.com/IIIA-ML/geoloc>, 2022. [Last accessed 30-May-2022].
- [2] Artificial Intelligence Research Institute (IIIA), CSIC. Datasets for image geolocation. <https://github.com/IIIA-ML/geoloc-data>, 2022. [Last accessed 30-May-2022].

- [3] Yannis Avrithis, Yannis Kalantidis, Giorgos Tolias, and Evangelos Spyrou. Retrieving landmark and non-landmark images from community photo collections. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 153–162, 2010.
- [4] Sara Barozzi, Jose Luis Fernandez Marquez, Amudha Ravi Shankar, Barbara Pernici, et al. Filtering images extracted from social media in the response phase of emergency events. In *16th Conference on Information Systems for Crisis Response and Management*, pages 1–12, 2019.
- [5] Clemens Havas, Bernd Resch, Chiara Francalanci, Barbara Pernici, Gabriele Scalia, Jose Luis Fernandez-Marquez, Tim Van Achte, Gunter Zeug, Maria Rosa (Rosy) Mondardini, Domenico Grandoni, Birgit Kirsch, Milan Kalas, Valerio Lorini, and Stefan Rüping. E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, 17(12):2766, 2017.
- [6] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.
- [7] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *European conference on computer vision*, pages 15–29. Springer, 2012.
- [8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [9] Luca Loria. Hierarchical classification model for content-based geolocation of outdoor images with visual explanations. Master’s thesis, Politecnico di Milano, 2021.
- [10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [12] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021.
- [13] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Computer Vision – ECCV 2018*, pages 575–592, Cham, 2018. Springer International Publishing.
- [14] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 47–56, 2008.
- [15] Amudha Ravi Shankar, Jose Luis Fernandez-Marquez, Barbara Pernici, Gabriele Scalia, Maria Rosa Mondardini, and Giovanna Di Marzo Serugendo. Crowd4ems: A crowdsourcing platform for gathering and geolocating social media content in disaster response. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:331–340, 2019.
- [16] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *European conference on computer vision*, pages 752–765. Springer, 2012.

- [17] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.
- [18] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - photo geolocation with convolutional neural networks. <http://arxiv.org/abs/1602.05314>.
- [19] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. <https://arxiv.org/abs/2203.11926>, 2022.

Analyzing the Reliability of Different Machine Radiomics Features Considering Various Segmentation Approaches in Lung Cancer CT Images

Maryam TAHMOORESI^{a,1}, Mohamed ABDEL-NASSER^{a,b}, and Domenec PUIG^a

^aDepartment of Computer Engineering and Mathematics, University Rovira i Virgili,
43007 Tarragona, Spain

^bElectrical Engineering Department, Aswan University, Aswan 81528, Egypt

Abstract. Cancer is generally defined as the uncontrollable increase of number of cells in the body. These cells might be formed anywhere in the body and spread to other parts of the body. Although the mortality rate of cancer is high, it is possible to decrease cancer cases by up to 30% to 50% through taking a healthy lifestyle and avoiding unhealthy habits. Imaging is one of the powerful technologies used for detecting and treating cancer at its early stages. Nowadays, scientists admit that medical images hold more information than their diagnosis, which is called a radiomics approach. Radiomics demonstrate that images comprise numerous quantitative features that are useful in predicting, detecting, and treating cancers in a personalized manner. While radiomics can extract numerous features, not all of them are useful. It should not be neglected that the outcome of data analysis is highly dependent on the selected features. There are different ways of finding the most reliable features. One possible way is to select all extracted features, analyze them, and find the most reproducible and reliable ones. Different statistical analysis metrics could analyze the features. To discover and introduce the most accurate metrics, in this paper, different statistical metrics used for measuring the stability and reproducibility of the features are investigated.

Keywords. CAD, reliability, radiomics, lung cancer

1. Introduction

According to WHO cancer is one of the main causes of death with around 10 million deaths in 2021. Although the mortality of cancer cases is high, cancer cases can be prevented from 30% to 50%, and the others diagnosed in the early stages and with the appropriate treatment can be controlled or be cured.

One of the powerful technologies that are used for early detection and treatment is imaging [1]. Medical images allow us to visualize the entire body that is not possible to see by the naked eye [2]. Among different image types, MRI, ultrasound, PET, and CT are used widely for early detection, finding the stage of cancer, morphology, density change, etc. [3]. In the past, medical images were used to diagnose the presence of tumors, and this diagnosis depended on the experience and knowledge of physicians, but after a

¹ Corresponding Author: maryam.tahmooresi@yahoo.com .

while, with the advancement of science and the advent of artificial intelligence, efforts to extract more accurate information images continued to reduce the dependence of results on human science, thus reducing time and error. In recent years, scientists have come to realize that medical images contain information more than they diagnose [3] and this is called a radiomics approach. Radiomics shows that images contain innumerable quantitative features which can be used for predictive, detective, prognostic, and treatment personalized of the cancers [4][5][6].

Although radiomics can extract thousands of features, it does not mean that they are all useful. In addition, it is necessary to consider that the result of data analysis methods strictly depends on the chosen features, and it can be affected by some of them, thus it might lead to achieving poor results. Therefore, finding the robustness features to make the model, is of utmost important [7].

There are different ways to find the most reliable features, one of these ways is to select all extracted features and analyze them to find the most reliable and reproducible feature [7]. There are different statistical analysis metrics to analyze the features. In this work, we investigate different statistical metrics used to measure the features' stability and reproducibility to introduce the most accurate ones.

According to the data that we have, ICC is the most useful metric, but it has some limitations mentioned in the next sections. For this reason, the main aim of this paper is to use ICC and other appropriate metrics and compare the results to find an alternative metric. Our aim is broken into three objectives that were followed. First, compare manual and semi-automatic segmentation to find out the more accurate one using ICC and Kruskal. The result of the experiments proved that semi-automatic is better than manual segmentation. Second, the comparison was repeated by Kruskal, but this time 2 trainer oncologists were eliminated to see whether it could improve the manual segmentation or not. This semi-automatic time segmentation shows better performance, too. Additionally, the feature categories were compared to determine which one had more reliable features. The results show that first order and NGTDM have the best results with 100%.

2. Related Works

Among different statistical metrics to evaluate the extracted features. Intra-class Correlation Coefficient is the one applied by most researchers for different reliability analysis types like test-retest, interrater, and interrater [8].

Baeßler, Weiss, and Dos Santos [9] worked on MRI to find the robustness features. Therefore, they chose different fruits/vegetables and scanned them by using FLAIR, T1W, and T2W with high and low resolutions. Later, the extracted features were used for test-retest and intraobserver and interobserver analysis., concordance correlation coefficient (CCC) was used for test-retest and, intraclass correlation coefficient (ICC) was implemented for interobserver and intraobserver. According to the achieved results, high-resolution FLAIR images showed the most reliable features, and they can be used for medical aims but in the case of T1W and T2W, it is necessary to take care to choose the features.

Lee et al. [10] focused on MRI scanning protocol parameters to find the effect of different parameters on radiomics features. They used some parameters like T1W, T2W, NEX, etc. with two scanners. ICC and COV were used to analyze the results of the test-retest scheme. The results show that scanning parameters and scanners which are used,

can affect the radiomics features and among these features, the ones with high ICC and CV can be considered reliable features to use.

Zwanenburg et al. [11] worked on robustness radiomics features for CT scan images by adding noise, translation, rotation, etc. as an alternative way for test-retest analysis. For this aim, they worked on two cancer datasets including non-small-cell lung cancer (NSCLC) and head-and-neck squamous cell carcinoma (HNSCC) to check the reproducibility of the extracted features for perturbation and test-retest and compared the results. ICC is the statistical metric that is used for the measurement and showed that this perturbation chain may use instead of test-retest.

Fiset et al. [12] worked on finding reliable radiomics features for cervical cancer and MRI is the image that they selected to work. They performed their analysis in three models including test-retest, diagnostic MRI and simulation MRI, and interobserver to find out which model can produce the most reproducible features. ICC showed that features of the test-retest chain are the most reliable and among the features' categories, shape features are the best ones.

As we mentioned before ICC is the most useful statistical metric, is used to find the robustness radiomics features, but this metric has some limitations. Here we review some papers that mentioned the ICC limitations.

Mehta et al. [13] worked on the dependency of ICC to subject distribution and sample size. They found out that convex distribution has less ICC than uniform distribution and even, less than concave distribution and this dependency is a problem to using the ICC results for reliability analysis. In the second step, they checked the effect of sample size. Thus, they used a fixed type of distribution and the findings proved that increasing the number of samples has an impact on ICC until for example $n=80$ and after that, there is no effect. They believe that, although most researchers use ICC for analysis, they should be aware of its conditions, usage, and limitations.

Pleil, Wallace, Stiegel, and Funk [14] studied articles for explaining the importance of repeat measures in biomonitoring research to assess variability and eventually calculating health risk. The aim is (1) to introduce the idea of creating measurements for biomarkers, (2) to review the records of using ICC (intra-class correlation coefficients) in health-based decision making, and (3) to examine the effectiveness of various methods in ICC calculation making. According to the result of ICC, they argue that ICC estimates' precision is highly influenced by the sum of samples, the number of repeat measures, and the special sample distribution.

Chen and Barnhart [15] worked on ICC and CCC and believe these are the most common metrics which are used for analyzing reliability. Not only do they consider the effects of subject and observer with repeated measurements, but also the effects of time on data. Because practically, it is not easy to gain the true replications. these two indices of the agreement for various combinations of fixed or random effects of time and observer are compared. ultimately, 2D-echocardiogram image data is used for illustrating the suggested methodology and comparing these 2 indices. In case, of repeated measurements, one needs to choose between these two indices, using a new concordance correlation coefficient is recommended.

According to the limitations of the ICC, we aim to use other metrics to find the most reliable features. ICC and other metrics are calculated for the extracted features. The achieved results will be compared to find the best metric. In the next section, we explain this process in detail.

3. Methodology

3.1. Dataset

The Cancer Imaging Archive (TCIA) including different datasets for various cancer types is used in this study. the non-small cell lung cancer (NSCLC) dataset was used which included 22 patients' CT images [16]. The dataset is comprised of manual and semi-automatic segmentation, where the manual segmentation was performed by 5 different radiation oncologists (3 experienced and 2 trainer oncologists). These 5 people also did the semi-automatic segmentation. In case of the need for any correction, they used in-house automatic segmentation tools and checking segmentation.

One patient among these 22, does not have tumor delineations and another one just has manual delineations. So, they were eliminated from our study and the rest of the 20 patients' images were used to have the same number of images for both segmentations.

3.2. Features Definitions and Extraction

By radiomics, thousands of quantitative features could be extracted, that can describe lesions, and they are divided into four main categories of shape, first-order, second-order, and higher-order features [7] [3]. We can extract these features directly or after applying any type of filter. Here each category is defined in short:

- Shape: It is defined as the main features that describe the ROI size and shape for example maximum diameters, surface area, volume, etc.
- First-order Features: The first-order features are normally based on a histogram and recount the spread and position of each voxel value without regard to kurtosis, skewness, uniformity, or other spatial relationships.
- Second-order Features: The second-order features which are generally called texture features, explain neighboring voxels' inter-relationship and it is categorized into the following subgroups: Gray-Level Size Zone Matrix (GLSZM), Gray-Level Co-occurrence Matrix (GLCM), Neighbourhood Gray-Tone Difference Matrix (NGTDM), Gray-Level Run Length Matrix (GLRLM), and Gray- Level Dependence Matrix (GLDM). Each of these subgroups contains different features.
- Higher-order Features: After the implementation of any filter or mathematical transform, higher-order features are achieved. For instance, any filter or transform such as Fractal analysis, wavelet transform, and Laplacian can be applied for bolding the details or finding the repetitive and non-repetitive patterns.

The application used for this study was a 3D slicer to extract the features, a wavelet was used also the whole features were 851. The purpose is to select all features and analyze them by using statistical metrics to obtain the most reliable and accurate features.

3.3. Impact of Image Segmentation

Finding the incorrect region of interest (ROI) might lead to poor results, as the features are extracted from this region [18], thus, Lesion delineation is one of the most significant challenges of radiomics. Most of the tumors do not show clear borders, so it is a

challenging task and even though delegating the responsibility to the experts can be helpful, as shown in Fig 1, the result of the borders which was shown by the three expert oncologists for the same lesion, were not the same.

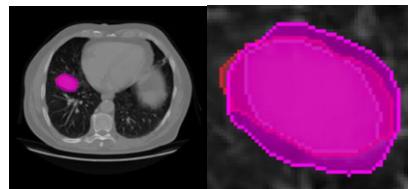


Figure 1.3 Manual segmentation.

On the other hand, to save energy, and time and most importantly decrease the mistake level and improve the performance, automatic and semi-automatic segmentation techniques were developed by improving technology. Thus, in the present paper, one of our experiments was to compare manual and semi-automatic segmentation to see whether the semi-automatic one is more accurate or not.

3.4. Evaluation Metrics and Statistical Analysis

As mentioned earlier, radiomics enables us to extract thousands of features but that does not guarantee the usefulness of all of them. Therefore, one feels the necessity of finding the most reproducible and reliable ones. To quantify the reproducibility of the features, Statistical metrics are implemented. Based on the type of data and the purpose, different metrics are available but in the present paper ICC and Kruskal-Wallis tests were used. Kruskal-Wallis Test was applied to see whether there is a statistically significant difference in 3 or more independent groups' medians. Kruskal-Wallis Test is a non-parametric version of the one-way ANOVA. Compared to the one-way ANOVA, in The Kruskal-Wallis test normality is not assumed in the data and it is not much sensitive to outliers, thus typically, if the normality assumption gets violated Kruskal-Wallis Test is used², which can be defined as follows:

$$H = (N - 1) \frac{\sum_{i=1}^g n_i (\bar{r}_i - \bar{r})^2}{\sum_{i=1}^g \sum_{j=1}^{n_i} (\bar{r}_{ij} - \bar{r})^2}$$

4. Results

4.1. Inter-rater reliability

851 features are extracted from our dataset by applying wavelet and they were divided into 9 groups (original, HHH, HHL, HLH, HLL, LHH, LHL, LLH, and LLL). Each group comprises six feature categories (first order, NGTDM, GLDM, GLCM, GLSZM, and GLRLM). In the original group, there is one more category of shape included.

² Zach, "Kruskal-Wallis Test: Definition, Formula, and Example," 2019. <https://www.statology.org/kruskal-wallis-test/>.

All features have five semi-automatic segmentation and 5 manual segmentations. To find out how many features are not affected by changing the interobserver, the first step was calculating Kruskal for both segmentations. The result is shown in Fig. 2.

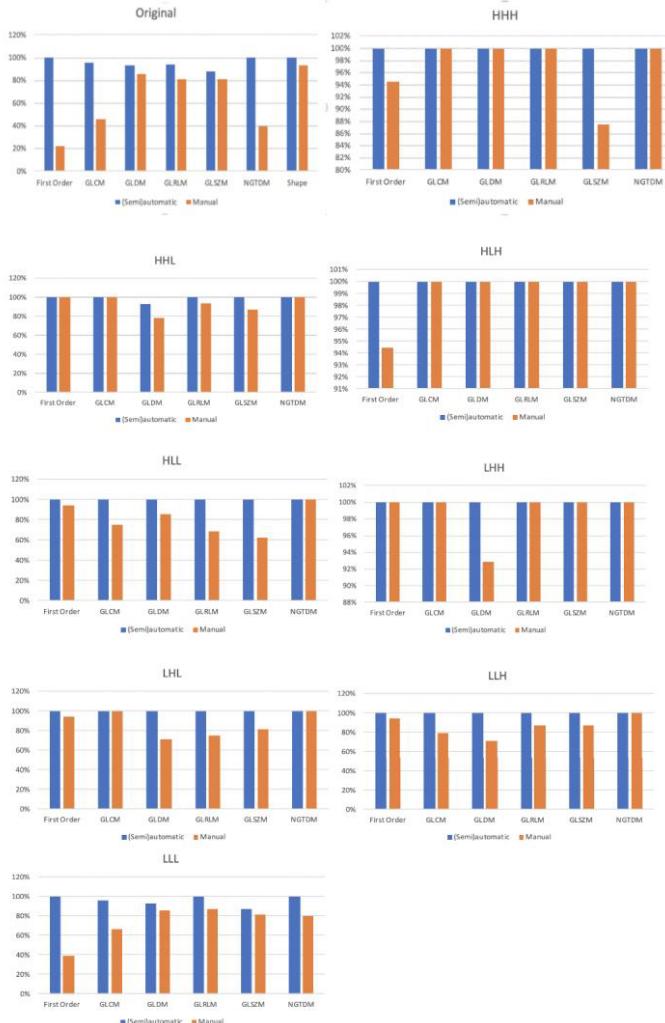


Figure 2. Comparison of semi-automatic and manual results.

As represented in Fig.2, the inter-rater reliability degree in semi-automatic segmentation and that of the manual segmentation is the same for some categories, but it is shown more in other categories.

The second step was comparing all categories for all groups to find out the group and category which has a more accurate interobserver degree. According to the attained results, NGTDM and first order are the ones with the most reliable features and different semi-automatic segmentation did not affect their features. Only the semi-automatic

segmentation has affected GLRLM's features in the original image type, but there is no effect after applying the filter.

As a second experiment, the ICC is calculated for the expressed features to check which of the segments has the most agreement. Table 1 shows in each group how many percent the manual segmentation showed better results than the semi-automatic segment.

Table 1. Manual segmentation's result (the percentage which manual has better results than semi-automatic)

	Original	HHH	HHL	HLH	LHH	LHL	HLL	LLH	LLL
First-order	5.5	25	28	0	22	16	0	11	18
GLDM	21	20	8	25	4	25	0	8	25
GLCM	25	57	14	21	14	21	21	14	21
GLRLM	18	18	12	25	12	0	37	18	12
GLSZM	6	25	25	12	31	25	37	31	25
NGTDM	20	40	0	40	0	40	0	20	0
SHAPE	0	-	-	-	-	-	-	-	-



Figure 3. Comparison of semi-automatic and manual segmentations' results (3 observers)

4.2. Inter-rater reliability

To calculate the Kruskal in the first experiment, we worked on the 10 segmentations (5 semi-automatic and 5 manual). In the second experiment, we plan on taking out the information of the two trainer oncologists from the calculations to see whether the results would be affected or not. As shown in Fig 3, like the first experiment, in this experiment again semi-automatic segmentation represents better results than manual segmentation. There are some cases where both have the same results but there are no features where manual segmentation shows better results.

5. Conclusion

In this paper, three objectives were followed. First, compare manual and semi-automatic segmentation to find the more accurate one. The result of the experiment proved that semi-automatic is better than manual segmentation. Second, the comparison was repeated, but this time two trainer oncologists were eliminated to see whether it could improve the manual segmentation or not. This semi-automatic time segmentation shows better performance, too. The feature categories were compared to determine which one had more reliable features. The results show that first order and NGTDM have the best result with 100%. One of the limitations we faced in this study is the implemented statistical metric. Kruskal is just applicable to accept or reject the null hypothesis, but it cannot show the degree of agreement between the observers. In future studies, we will work on this issue.

References

- [1] Liu R, Elhalawani H, Radwan MA, Elgohari B, Court L, Zhu H, Fuller CD., Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer, *Clin. Transl. Radiat. Oncol.*, vol. 21, pp. 11–18, 2020, doi: 10.1016/j.ctro.2019.11.005.
- [2] Mondal SB, O'Brien CM, Bishop K, Fields RC, Margenthaler JA, Achilefu S. Repurposing molecular imaging and sensing for cancer image-guided surgery, *J. Nucl. Med.*, vol. 61, no. 8, pp. 1113–1122, 2020, doi: 10.2967/jnumed.118.220426.
- [3] Jie T, Di D, Zhenyu L, Jingwei W. Radiomics and Its Clinical Application: Artificial Intelligence and Medical Big Data. Elsevier Science, 2021.
- [4] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data, *Radiology*, vol. 278, no. 2, pp. 563–577, 2016, doi: 10.1148/radiol.2015151169.
- [5] Yang F, Dogan N, Stoyanova R, Ford JC. Evaluation of radiomic texture feature error due to MRI acquisition and reconstruction: A simulation study utilizing ground truth, *Phys. Medica*, vol. 50, no. December 2017, pp. 26–36, 2018, doi: 10.1016/j.ejmp.2018.05.017.
- [6] Lambin P, Leijenaar RTH, Deist TM, Peerlings J, De Jong EEC, Van Timmeren J, Sanduleanu S, Larue RTHM, Even AJG, Jochems A, Van Wijk Y, Woodruff H, Van Soest J, Lustberg T, Roelofs E, Van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: The bridge between medical imaging and personalized medicine, *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017, doi: 10.1038/nrclinonc.2017.141.
- [7] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M .Radiomics: the facts and the challenges of image analysis, *Eur. Radiol. Exp.*, vol. 2, no. 1, 2018, doi: 10.1186/s41747-018-0068-z.
- [8] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research, *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.
- [9] Baeßler B, Weiss K, Santos DPD. “Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study,” *Invest. Radiol.*, vol. 54, no. 4, pp. 221–228, 2019, doi: 10.1097/RLI.0000000000000530.

- [10] Lee, J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, Yang J, Followill D, Ger R, MacKin D, Court LE. Radiomics feature robustness as measured using an MRI phantom, *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-021-83593-3.
- [11] Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, Löck S. Assessing robustness of radiomic features by image perturbation, *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019, doi: 10.1038/s41598-018-36938-4.
- [12] Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL., Milosevic M, Fyles A, Traverso A, Jaffray D, Metser U, Xie J, Han K. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, *Radiother. Oncol.*, vol. 135, pp. 107–114, 2019, doi: 10.1016/j.radonc.2019.03.001.
- [13] Mehta S, Bastero-Caballero RF, Sun Y, Zhu R, Murphy DK, Hardas B, Koch G. Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies, *Stat. Med.*, vol. 37, no. 18, pp. 2734–2752, 2018, doi: 10.1002/sim.7679.
- [14] Pleil JD, Wallace MAG, Stiegel MA, Funk WE. Human biomarker interpretation: the importance of intra-class correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates, *J. Toxicol. Environ. Health. - Part B Crit. Rev.*, vol. 21, no. 3, pp. 161–180, 2018, doi: 10.1080/10937404.2018.1490128.
- [15] Chen CC, Barnhart HX. Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures, *Comput. Stat. Data Anal.*, vol. 60, no. 1, pp. 132–145, 2013, doi: 10.1016/j.csda.2012.11.004.
- [16] Aerts HJWL, Wee L, Rios Velazquez E, Leijenaar RTH, Parmar C, Grossmann P, Lambin P. NSCLC-Radiomics[Dataset], 2019. <https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI> (accessed Dec. 16, 2021).
- [17] Li R, Xing L, Napel S, Rubin DL. Radiomics and radiogenomics: technical basis and clinical applications. 2019.

Transformer-Based Radiomics for Predicting Breast Tumor Malignancy Score in Ultrasonography

Mohamed A. HASSANIEN^{a,1}, Vivek KUMAR SINGH^c, Domenec PUIG^a and Mohamed ABDEL-NASSER^{a,b}

^a*Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili,
43007 Tarragona, Spain*

^b*Electrical Engineering Department; Aswan University, Aswan, Egypt*

^c*Queen's University Belfast, United Kingdom*

Abstract. Breast cancer must be detected early to reduce the mortality rate. Ultrasound images can make it easier for the clinician to diagnose cases of dense breasts. This study presents a deep vision transformer-based approach for predicting breast cancer malignancy scores from ultrasound images. In particular, various state-of-the-art deep vision transformers such as BEiT, CaiT, Swin, XCiT, and ViFormer are adapted and trained to extract robust radiomics to classify breast tumors in ultrasound images as benign or malignant. The best-performing model is used to predict the malignancy score of each input ultrasound image. Experimental results revealed that the proposed approach achieves promising results for the detection of malignant tumors of the breast on ultrasound images.

Keywords. Radiomics, Breast cancer, Ultrasound imaging, CAD systems, Vision transformers

1. Introduction

Mammography is the most commonly used imaging technique to detect the early stages of breast cancer. Breast ultrasonography, however, substitutes mammography in the case of dense breasts as it can not penetrate through the tissue. In pregnant women, breast ultrasound is a viable alternative to mammography to prevent the use of radiation that can harm the fetus [1]. Computer-aided diagnostic (CAD) solutions help clinicians free themselves from processing multiple breast images of a patient, thereby improving the quality of clinical diagnostics [2,3,4]. The leading steps of a CAD system for classifying breast tumors with ultrasound images include region of interest (ROI) detection, tumor segmentation, feature extraction and selection, and machine learning-based classification model development.

Nevertheless, the accuracy of such CAD systems may be restricted due to the low signal-to-noise ratio (SNR) and the presence of artifacts such as speckle noise and shadows in the breast ultrasound images. Inadequate contact between the probe and the skin surface

¹Corresponding Author; E-mail: mohamed.abdelhameedhassanien@estudiants.urv.cat.

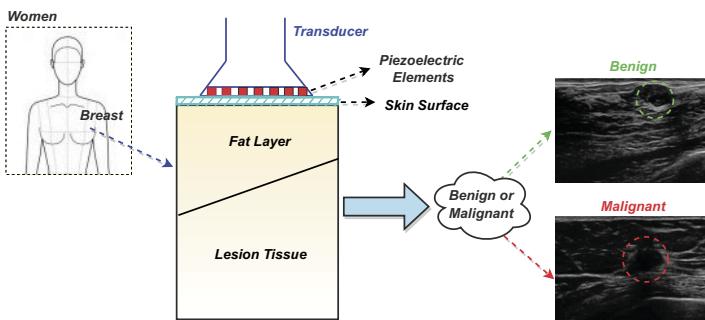


Figure 1. General description of breast tumor detection with breast ultrasound (BUS). Green and red dotted circles highlight the benign and malignant breast tumors respectively.

can cause a shadowing effect. These artifacts infer a clinical diagnosis, as it is difficult for a sonographer accurately acquire the appropriate breast tumor information.

Figure 1 shows a general description of breast tumor detection with breast ultrasound (BUS). Transducers with piezoelectric elements (128 or 192) generate sound waves reflected by the breast tissue to produce echoes. The ultrasound examiner or user must place the transducer correctly on the skin surface at the appropriate pressure to avoid the imaging artifact shadowing effects. The early layers of breast tissue contain subcutaneous fat and later depth approach to the lesion region. When an ultrasound scan is performed, the resulting B-mode image is examined by sonographers to see if the detected tumors are benign or malignant. Breast tumor is commonly recognized as a hypoechoic region than white breast tissue or light gray fat. The boundaries of most benign breast tumors are well defined and are round or oval in shape. In contrast, malignant tumor boundaries are often irregular and poorly defined, with lobules.

In the last two decades, many CAD systems have been proposed for breast cancer diagnosis. For instance, Shia et al. [5] used a deep residual network model to extract texture features of breast ultrasound images and then trained a support vector machine (SVM) to classify breast tumors as benign or malignant. The authors used an ultrasound dataset containing 302 benign and 241 malignant cases. The recommended method yielded a sensitivity score of 94.34% and a specificity score of 93.22%. In [6], Mishra et al. proposed a machine learning-based radiomics method classifying breast tumors with ultrasound. The authors extracted several hand-crafted features from the ultrasound images and later used recursive feature elimination to select the best set of those features. They also used synthetic minority oversampling techniques to address the problem of class imbalances. The employed texture features were Hu-moments based shape features, tumor shape features (area, convex area, eccentricity, solidity, EquivDiameter, extent, major axis length, minor axis length, orientation, and perimeter), histogram oriented gradients (HOG), and 13 grey-level co-occurrence matrix (GLCM). The authors employed an ultrasound dataset that included 437 benign, 210 malignant, and 133 normal cases in their experiments. They obtained results for accuracy and area under the curve (AUC) of 97.4% and 97%, respectively.

Zhuang et al. [7] applied the fuzzy enhancement, bilateral filtering, and image morphology operations on breast ultrasound images and the corresponding masks (binary image contains the segmented tumor). The authors concatenated various combinations of

the original and processed images, with each combination comprising three images (i.e., RGB channel fusion). The fused images were fed into pre-trained CNN models in order to extract feature vectors, which were subsequently merged using the adaptive spatial feature fusion approach. Finally, an artificial neural network (ANN) was used for the final classification. A dataset of 1328 breast ultrasound images was utilized to train and test this approach, which yielded an accuracy of 95.48%. Cui et al. [8] proposed two enhanced combined-tumoral region modules to gradually enhance the combined-tumoral features. Besides, the authors proposed a three-stream module for extracting and combining intratumoral, peritumoral, and combined tumor area features. The author used the channel attention module to adaptively combine the features of the three regions. The proposed method achieved a precision of 94.50% using the UDIAT dataset. Yu et al. [9] extracted discriminative regions from the input ultrasound images: the inner region, the marginal zone, and the posterior echo region of the lesion image. Then, they used an Inception-V3 pre-trained CNN model to extract texture features from the three regions and the whole image, followed by a principal components analysis (PCA) to reduce the dimensionality of the extracted features and an ensemble learning classifier for classifying the input image as benign or malignant. With a dataset of 479 cases, they achieved an accuracy of 85%.

Recently, several transformer-based methods have advanced exponentially for medical imaging tasks. Transformers have proven their capability to capture long-range dependencies and learn better relevant feature representations to be an alternative to convolutional neural networks. In the ultrasound domain, only limited research has been conducted to measure the effectiveness of transformer methods for classifying breast tumors in challenging conditions. This paper presents a deep learning-based radiomics approach for detecting breast tumor malignancy. Various self-attention based deep vision transformer architectures are adapted and trained to extract robust radiomics to classify breast cancers as benign or malignant, and predict the malignancy score of each input ultrasound image. Based on the transfer learning theory, the pretrained vision transformer network and parameters can be applied to this study's target breast ultrasound dataset by fine-tuning the selected vision transformer networks. Furthermore, we investigate the feasibility of incorporating the malignancy scores of the top-performing transformer model to enhance detection accuracy.

The rest of this paper is organized as follows. Section 2 details the proposed methodology. Section 3 contains experimental results and discussions.. We conclude our finding and suggest some future lines of research in Section 4.

2. Methodology

2.1. Breast cancer malignancy score prediction

Figure 2 presents the proposed radiomics approach for detecting breast tumor malignancy. A set of deep vision transformers are used to extract robust radiomics and classify breast tumors as benign or malignant. Various data augmentation techniques are used to increase the number of ultrasound images for training. The best model is identified based on different evaluation metrics. The extracted radiomics are used to compute the malignancy scores of breast cancer from the input ultrasound image.

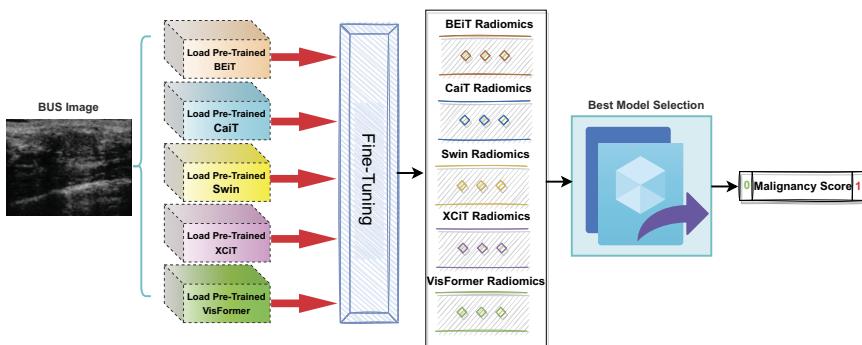


Figure 2. The pipeline of the proposed approach for breast tumor malignancy prediction in breast ultrasound images.

2.2. Input ultrasound images

The ultrasound images were taken from the UDIAT Diagnostic Centre of Sabadell (Spain) [10], [11]. It has a total of 163 breast ultrasound images extracted from 263 patients using a Siemens ACUSON Sequoia C512 system. It comprises two classes benign and malignant which have 110 and 53 breast ultrasound images, respectively. The annotation for each image of the lesion area was available.

2.3. Data augmentation

We applied several data augmentation techniques such as horizontal and vertical flipping with the probability of 0.5, scaling with 0.5, contrast limited adaptive histogram equalization (CLAHE), and rotation of 30 degrees which provides additional feature variability to the network. It is worth noting that all the trained transformer-based networks used the same data augmentation techniques.

2.4. Constructing transformer-based radiomics

In this study, five deep transformer models have been employed, namely, BEiT [12], CaiT [13], Swin [14], XCiT [15], and Visformer [16], for extracting deep learning-based radiomics for breast cancer malignancy prediction. Below, we briefly explain the architecture of each network.

- BEiT [12] is a self-supervised vision representation model from Image Transformers. In BEiT's pre-training, each Image has two views: image patches (16x16 pixels) and visual tokens (i.e., discrete tokens). Tokenization of the source image into visual tokens occurs first. The backbone Transformer is then fed with some randomly masked image patches. The pre-training goal is to use the corrupted image patches to recover the original visual tokens. BEiT's model parameters are fine-tuned on downstream tasks by appending task layers to the pre-trained encoder after it has been pre-trained. The implementation of the BEiT transformer is available at <https://github.com/microsoft/unilm/tree/master/beit>.

- CaiT [13] stands for class attention in image transformers. It is a way of enhancing the training of more profound architecture compared to other image transformer approaches. A LayerScale technique is employed in CaiT, where a learnable diagonal matrix is attached to the output of each residual block, in which some elements are initialized with values close to zero. The training dynamic is enhanced by attaching this basic layer after each residual block, allowing for the training of deeper image transformers. As a result, models whose performance does not saturate early with increased depth are produced. The implementation of the CaiT transformer is available at <https://github.com/facebookresearch/deit>.
- The Swin transformer [14] is a general-purpose transformer backbone that creates hierarchical feature maps with linear computing complexity concerning image size. Swin transformer creates a hierarchical representation by fusing neighboring patches (i.e., shifting windowing) in deeper transformer layers, starting with small patches. The shifted windowing approach enhances efficiency by limiting self-attention calculation to non-overlapping local windows while enabling cross-window connectivity. The hierarchical architecture of the Swin transformer facilitates the modeling of different scales. The implementation of the Swin transformer is available at <https://github.com/microsoft/Swin-Transformer>.
- XCiT (cross-covariance image transformer) [15] coalesces convolutional architecture scalability with classical transformer accuracy. It comprises a transposed variant of self-attention that interacts using the cross-covariance matrix between keys and queries rather than tokens. The resulting cross-covariance attention is linear in terms of token complexity and can swiftly analyze high-resolution images. Regardless of the number of tokens, XCiT attends to a predetermined number of channels. As a result, XCiT is far more resistant to variations in image resolution during testing and hence better suited to processing variable-size images. The implementation of XCiT is available at <https://github.com/facebookresearch/xcit>.
- VisFormer (vision-friendly transformer) [16] improves visual identification by switching from a transformer-based model to a convolution-based one. VisFormer uses a gradual transition technique to bridge the gap between transformer-based and convolution-based models, revealing the features of the designs in both. It dissected the gap between these models and devised an eight-step transition process to connect DeiT-S and ResNet-50. The implementation of VisFormer is available at <https://github.com/danczs/Visformer>.

2.5. Implementation details

We resized the original BUS image resolution to the 224×224 pixels and calculated the mean and standard deviation to normalize the dataset. An Adam optimizer was used with an initial learning rate of 0.0001 and selected the default value of β_1 and β_2 to 0.9 and 0.999, respectively. All the transformer-based networks were trained at 40 epochs with the mini-batch of four samples and saved the best checkpoint of highest classification accuracy on validation to evaluate the performance on the independent test set. We used the cross-entropy loss function to minimize the error during network training. It should be noted that all the networks utilized the same hyperparameter setting to train and evaluate the classification performance. We divided the dataset into the three subsets of training,

validation, and test with ratios of 70%, 10%, and 20%, respectively. We trained and evaluated all the transformer-based models on PyTorch with NVIDIA GeForce GTX 1070Ti GPU of 8GB RAM.

2.6. Evaluation metrics

In this study, the performance of the proposed approach has been assessed using different evaluation metrics, namely accuracy, precision, recall, and F1-score. These metrics can be defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F1-score} = TP / (TP + 0.5(FP + FN)) \quad (4)$$

where TP represents the number of malignant cases correctly classified as malignant; TN represents the number of benign cases correctly classified as benign; FP represents the number of benign cases incorrectly classified as malignant, and FN represents the number of malignant cases incorrectly classified as benign.

3. Experimental Results and Discussion

Table 1 presents the breast ultrasound classification results of the deep radiomics for the five state-of-the-art transformer-based deep learning methods. It includes the BEiT [12], CaiT [13], Swin [14], XCiT [15], and VisFormer [16]. Figure 3, and 4 demonstrate the confusion matrix of each methods on both the test and validation sets. The test set contains a total of 22 and 11 samples for benign (0) and malignant (1) classes, respectively. In the test set, Swin and XCiT showed comparable performance and achieved an accuracy of 87.88%. The Swin transformer offers the benefits of a small BUS patch and exponentially increases its size by fusing to maintain scale-invariant properties. This mechanism helps capture lesions of various sizes in BUS images and provides robust feature representation to distinguish between benign and malignant tissue patterns. XCiT, on the other hand, provides an efficient self-attention mechanism for BUS features, acting through functional channels instead of tokens, improving classification performance. Both the Swin and XCiT models achieved a precision rate, recall, and F1 score of over 85%. However, due to the increased complexity (i.e., 300M parameter), BEiT has achieved an accuracy of 66.66% less accurately than existing methods. It failed to learn the pattern of benign and malignant classes that require more train samples to effectively achieve better results. CaiT provided the second-highest result, 6% lower than the Swin and XCiT methods. In addition, VisFormer has reached an accuracy of 75.76%.

We calculated each method's computational complexity by measuring the trainable parameters. As one can see, the BEiT method achieved the highest number of parameters (300 million) than existing methods that require additional breast lesion ultrasound samples to improve the classification performance. In turn, Swin, XCiT, VisFormer, and CaiT

Table 1. State-of-the-art Transformer based models comparison on BUS dataset. The best results are highlighted in bold.

Methods	Evaluation Metrics				
	Accuracy	Precision	Recall	F1-Score	Parameters (M)
BEiT-Radiomics	66.66	100.00	66.66	80.00	300
CaiT-Radiomics	81.82	81.82	90.00	85.71	46.5
XCiT-Radiomics	87.88	86.36	95.00	90.48	188.16
VisFormer-Radiomics	75.76	95.46	75.00	84.00	39.45
Swin-Radiomics	87.88	90.91	90.91	90.91	195

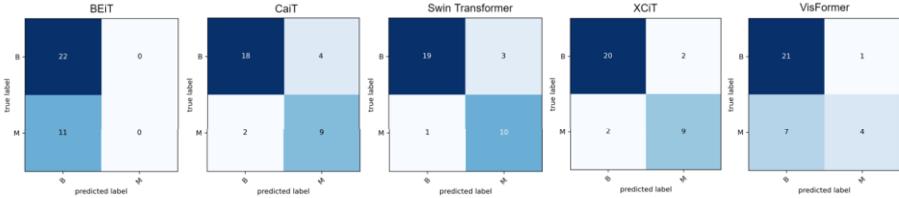


Figure 3. Confusion matrix on test set.

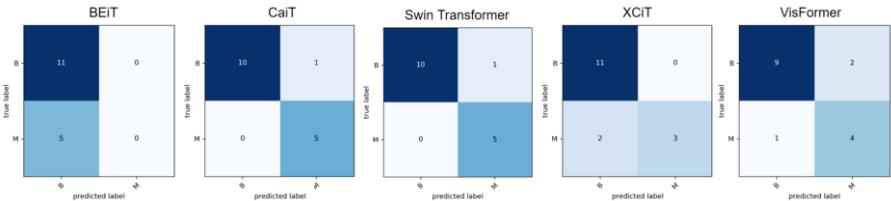


Figure 4. Confusion matrix on val set.

utilized the 194M, 188M, 39M, and 46M trainable parameters respectively. Conclusively, we have found that Swin transformer efficiently extract the radiomics features from BUS ultrasound and achieved state-of-the-art results compared to recently published works.

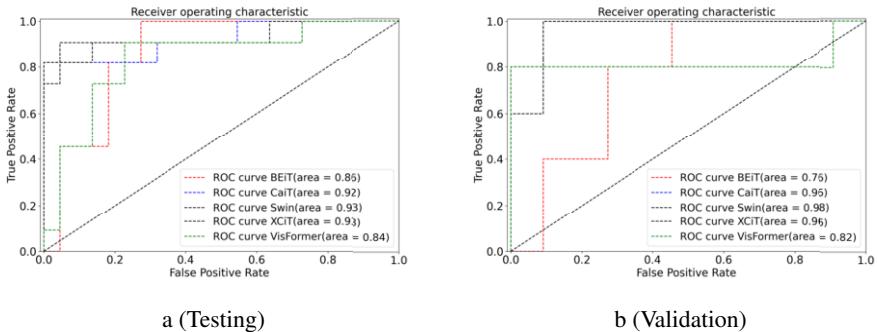


Figure 5. ROC curves and AUC values of the radiomics of test, and validation.

Figure 5 shows the receiver operating characteristics (ROC) curves and AUC values of all the examined models. We computed the ROC curve for both evaluated testing and

validation sets. On the validation samples, the Swin transformer has achieved the highest AUC score of 0.98%. However, CaiT and XCiT obtained identical results of 0.96%. The BEiT has only yielded the AUC value of 0.76% which was the lowest of all the transformer-based compared methods. However, on the testing set, the Swin transformer and XCiT have obtained the equal highest AUC score of 0.93% than other existing deep models. At the same time, CaiT obtained the AUC values of 0.92%. The VisFormer yielded the lowest AUC score of 0.84% compared to other methods, while BEiT only attained the 0.86%.

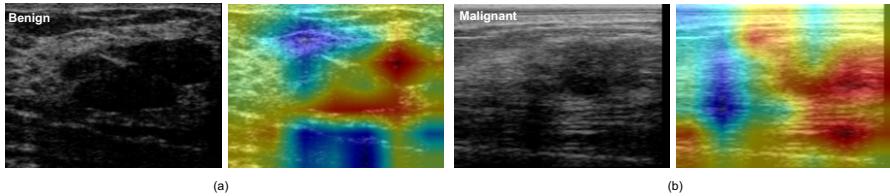


Figure 6. Illustration of Gradcam visualization generated with Swin transformer for two examples of breast ultrasound images. The higher the intensity of the red color, the more attention the model pays to the area of interest.

Figure 6 shows the Gradcam visualization of the best results obtained by the Swin transformer on an ultrasound image of the breast tissues. The core idea was to investigate how the model filters capture the targeted lesion region underlying several noises presented in ultrasound. We found that the model precisely captured the hypoechoic lesion areas by paying more attention to the dense structure of the model. The two different examples presented from benign and malignant classes have different textural patterns and imaging characteristics. For the benign cases, the lesion presented in ultrasound has increased neighboring artifacts that are surrounded by shadows and speckle noises. Due to the great feature representation of the Swin transformer, it is noticeable that it can efficiently highlight the relevant lesion features (red) and ignores the noisy background regions (blue). However, it shows similar characteristics for malignant samples were correctly identified the breast lesion pixels and focused less on neighboring artifacts.

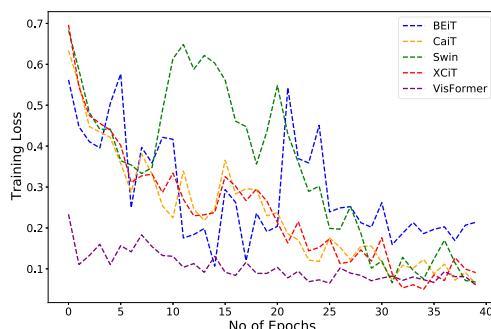


Figure 7. Convergence of each transformer-based methods.

We also investigated the convergence of individually trained deep models such as BEiT,

CaiT, Swin, XCiT and VisFormer. Figure 7 shows the loss convergence of all five existing methods. We observed that all the methods has trained well and training error were minimized at epoch 40.

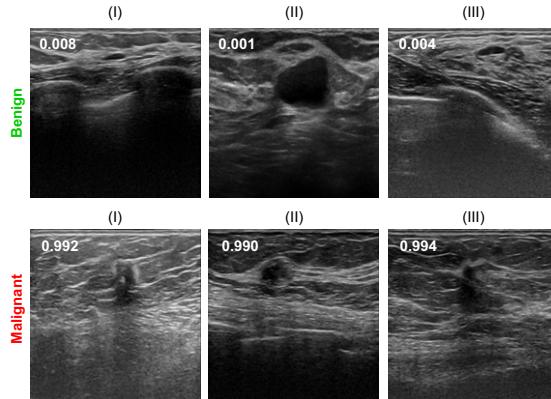


Figure 8. Illustration of malignancy scores (benign and malignant classes) achieved by the highest scores achieved by Swin Transformer. Three examples from each class, score are pasted on the image. High scores above 0.5 correspond to malignant otherwise benign.

Figure 8 shows the malignancy score examples from the test set obtained through the Swin transformer network. From the visual inspection, both benign and malignant class samples show distinct textural patterns. Malignant tumors have an ambiguous boundary, while benign-class tumors contain smooth structures with hypoechoic regions. The scores above 0.5 are malignant. The three malignant examples achieved very high scores of more than 0.99. In turn, the benign class samples obtained the lowest malignancy score. With efficient feature representation, the Swin transformer precisely classified the presented six samples into their classes.

4. Conclusions

A deep vision transformer-based strategy for predicting breast cancer malignancy scores from ultrasound pictures was proposed in this research. BEiT, CaiT, Swin, XCiT, and VisFormer are among the deep vision transformers that have been adopted and trained to extract robust radiomics for categorizing benign and malignant breast cancers in ultrasound pictures. For each input ultrasound image, the highest-performing model is used to forecast the malignancy score. The experimental results demonstrated that the Swin-based radiomics yielded the best classification results. Both the Swin and XCiT models achieved a precision rate, recall, and F1-score of over 90%. The future work will include analyzing other deep learning methods and investigating the use of different multi-model aggregation techniques for enhancing the classification results.

Acknowledgement

The Spanish Government partly supported this research through Project PID2019-105789RB-I00.

References

- [1] Carol H Lee, D David Dershaw, Daniel Kopans, Phil Evans, Barbara Monsees, Debra Monticciolo, R James Brenner, Lawrence Bassett, Wendie Berg, Stephen Feig, et al. Breast cancer screening with imaging: recommendations from the society of breast imaging and the acr on the use of mammography, breast mri, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. *Journal of the American college of radiology*, 7(1):18–27, 2010.
- [2] Vivek Kumar Singh, Mohamed Abdel-Nasser, Farhan Akram, Hatem A Rashwan, Md Mostafa Kamal Sarker, Nidhi Pandey, Santiago Romani, and Domenec Puig. Breast tumor segmentation in ultrasound images using contextual-information-aware deep adversarial learning framework. *Expert Systems with Applications*, 162:113870, 2020.
- [3] Mohamed Abdel-Nasser, Jaime Melendez, Antonio Moreno, and Domenec Puig. The impact of pixel resolution, integration scale, preprocessing, and feature normalization on texture analysis for mass classification in mammograms. *International Journal of Optics*, 2016.
- [4] M Abdel-Nasser, A Moreno, and D Puig. Temporal mammogram image registration using optimized curvilinear coordinates. *Computer Methods and Programs in Biomedicine*, 127:1–14, 2016.
- [5] Wei-Chung Shia and Dar-Ren Chen. Classification of malignant tumors in breast ultrasound using a pretrained deep residual network model and support vector machine. *Computerized Medical Imaging and Graphics*, 87:101829, 2021.
- [6] Arnab K Mishra, Pinki Roy, Sivaji Bandyopadhyay, and Sujit K Das. Breast ultrasound tumour classification: A machine learning—radiomics based approach. *Expert Systems*, 38(7):e12713, 2021.
- [7] Zhemin Zhuang, Zengbiao Yang, Alex Noel Joseph Raj, Chuliang Wei, Pengcheng Jin, and Shuxin Zhuang. Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. *Computer methods and programs in biomedicine*, 208:106221, 2021.
- [8] Wenju Cui, Yunsong Peng, Gang Yuan, Weiwei Cao, Yuzhu Cao, Zhengda Lu, Xinye Ni, Zhuangzhi Yan, and Jian Zheng. Fmrnet: A fused network of multiple tumoral regions for breast tumor classification with ultrasound images. *Medical Physics*, 49(1):144–157, 2022.
- [9] Hailong Yu, Hang Sun, Jing Li, Liying Shi, Nan Bao, Hong Li, Wei Qian, and Shi Zhou. Effective diagnostic model construction based on discriminative breast ultrasound image regions using deep feature extraction. *Medical Physics*, 48(6):2920–2928, 2021.
- [10] Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Martí. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017.
- [11] Mohamed Abdel-Nasser, Domenec Puig, Antonio Moreno, Adel Saleh, Joan Martí, Luis Martin, and Anna Magarolas. Breast tissue characterization in x-ray and ultrasound images using fuzzy local directional patterns and support vector machines. In *VISAPP (1)*, pages 387–394, 2015.
- [12] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [13] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.
- [14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [15] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34, 2021.
- [16] Zhensu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021.

EDBNet: Efficient Dual-Decoder Boosted Network for Eye Retinal Exudates Segmentation

Mohammed Yousef Salem ALI^{a,1}, Mohamed ABDEL-NASSER^{a,b}, Aida VALLS^a,
Marc BAGET^c and Mohammed JABREEL^a

^aITAKA, Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

^bDepartment of Electrical Engineering, Aswan University, 81528 Aswan, Egypt

^cIISPV, Hospital Universitari Sant Joan de Reus, Spain

Abstract. Diabetic retinopathy (DR) is one of the most common causes of vision loss or blindness globally. Early detection of retinal eye lesions like hard exudates, soft exudates, microaneurysms, and hemorrhages is crucial to detect DR in a human eye. Therefore, accurate segmentation of lesions from eye fundus images is essential to develop efficient automated DR detection systems. This paper presented a novel hard and soft exudates lesions segmentation method called Efficient Dual-Decoder Boosted Network (EDBNet). EDBNet is composed of the following main components: 1) pre-trained ImageNet ResNet50 encoder with Atrous Spatial Pyramid Pooling (ASPP), 2) UNet decoder block with Gated Skip Connections mechanism to enhance capture more details of fundus images, 3) dual-decoder boosted to improve the performance segmentation of retinal lesion in the eye fundus images, and fusion outputs of the dual-decoder boosted to generate enhanced exudates segmentation. The effectiveness of the proposed framework is assessed on the IDRiD publicly dataset in terms of accuracy, Area Under Precision-Recall (AUPR), IOU, and Dice metrics. EDBNet obtains 99.8, 74.4, 78.0, and 87.6% of soft exudates, respectively. For hard exudates, EDBNet achieves 99.5, 85.3, 80.3, and 89.1%, respectively. The experimental results also demonstrate that EDBNet outperforms many state-of-the-art methods.

Keywords. Fundus Images, Exudates, Lesion Segmentation, Deep Learning, Diabetic Retinopathy

1. Introduction

Diabetic retinopathy (DR) is a disease that highly affects human eye vision. The early detection and treatment of DR are essential to prevent total vision loss [1]. DR produces different retinal lesions called microaneurysm (MA), haemorrhage (HE), hard exudate (EX), and soft exudate (SE). Ophthalmologists inspect eye fundus images to detect the signs of such lesions. However, the complex structure of lesions, various sizes, differences in brightness and the inter-class similarity with other fundus tissues add more dif-

¹Corresponding Author: Mohammed Yousef Salem Ali. E-mail: horbio10@gmail.com

ficulties to the manual analysis of many fundus images. In addition, manual detection of the tiny lesions is extensive and consumes both time and effort.

Modern computer-aided diagnosis (CAD) systems can analyze medical images and provide a diagnosis as accurate as ophthalmologists with many years of experience [2]. Deep learning (DL) technologies have become the cornerstone of several modern CAD systems. Several DL-based automated systems have recently been proposed for segmenting eye retinal lesions. Most of them use Convolutional Neural Networks (CNNs) like UNet [3] to automatically learn representative and high-level features from the input fundus images to achieve accurate segmentation. For instance, The authors of [4] proposed CARNet for multi-lesion segmentation. CARNet feeds the whole image and patch image into ResNet50 and ResNet101 networks, respectively, using a single attention refinement decoder. They used IDRiD, E-ophtha and DDR datasets for evaluation. EAD-Net [5] presented a CNN-based system divided into an encoder module, dual attention module, and decoder module. They evaluated their work on two datasets: the E_ophtha_EX dataset for exudates and the IDRiD dataset for four kinds of lesions. In [6] authors developed a weakly-supervised framework for fundus lesion segmentation using grayscale and morphological features of lesions and a deep neural network with an attention mechanism and residual module. They evaluated their system by 1485 images extracted from the Messidor dataset and labelled by them. Paper [7] introduced scale-aware attention with different backbones to re-weight multi-scale features of decoders dynamically, and they evaluated it on IDRiD, E-ophtha and DDR datasets.

Although there are many advantages offered by deep learning techniques, especially those based on the UNet models, there is still a problem of dealing with tiny lesions like retinal exudates, so we need to develop a boosted mechanism for dealing with them. This paper proposes an accurate eye retinal exudates segmentation model called EDBNet, that takes as an input a fundus image and produces its corresponding lesion mask. The novelties of the proposed DL model are the following:

1. EDBNet uses dual decoders to boost the performance and produce an accurate mask of the input image. The main unit of each decoder is the Gated Skip Connection (GSC) network [8]. The reason behind using the GSC is the ability to focus on the most valuable feature from the decoder based on the features from the previous level.
2. The first decoder receives the skip connections from the encoder, where the second decoder takes the output of the first one as skip connections in a cascading manner. Hence, although we only use two decoders in this work, EDBNet can generalize to any level of cascaded boosted decoders. Such cascading and composing of multiple layers gives the network the ability to learn representations of data with multiple levels of abstraction [9].
3. The final output is obtained by fusing the output of the two decoders, which can be seen as a kind of online ensembling technique. We mean by online that the fusion operation is done in both the training and the inference phases.

We conducted extensive experiments on the well-known publicly available IDRiD dataset. We achieved competitive performance in one kind of lesion compared with the state-of-the-art systems in the IDRiD challenge and outperformed them in another.

The remainder of this paper is organized as follows. Section 2 explains the proposed Eye Retinal Exudates Segmentation system. Then, section 3 presents the experiments and results. Finally, the conclusion and future work are provided in section 4.

2. EDBNet

This section describes the proposed model, EDBNet. As shown in Figure 1, it is composed by the following parts: **Backbone** (colored in blue), also known as encoder layer, which aims to encode the input image and produce feature maps at multiple levels of scales. **Neck** (colored in purple), which is an Atrous Spatial Pyramid Pooling (ASPP) layer that helps to extract high-resolution features. **Head**, also known as decoder layer, a boosted dual decoder based on Gated-Skip connections network [8] followed by an output layer. We explain in detail each part in the following subsections.

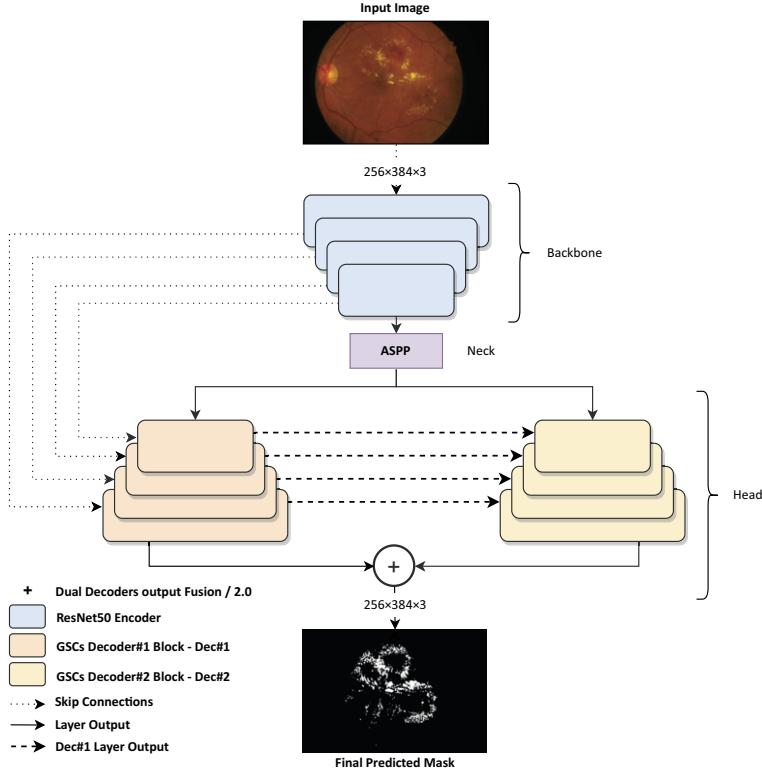


Figure 1. Structure of the EDBNet framework for Eye Retinal Exudates Segmentation.

2.1. Backbone: The Encoder Layer

We use a ResNet50 [10] encoder that was pre-trained on the ImageNet dataset as a backbone for our model. We selected this model because ResNet is the state-of-the-art backbone for many computer vision tasks [4,11].

The main goal of the backbone in our proposed model is to encode the input eye fundus image and extract abstracted and meaningful features at different levels of scales. The key advantage of ResNet50 is the residual connection which is used to add the output from an earlier layer to a later layer which helps to avoid the gradient vanishing problem.

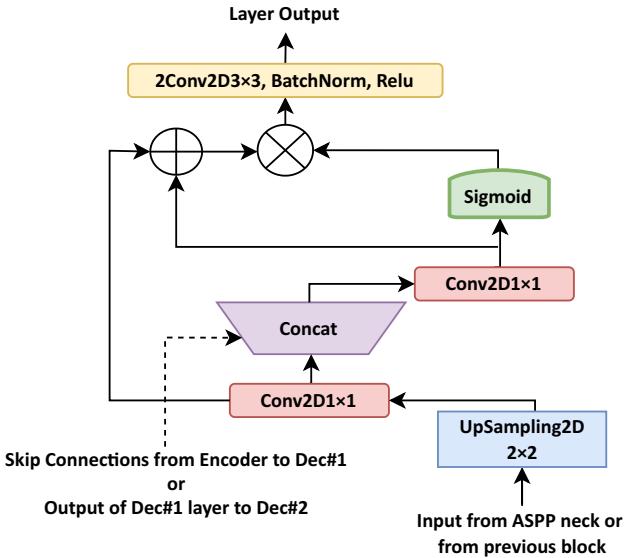


Figure 2. Gated Skip Connections network.

2.2. Neck: ASPP

To help extract high-resolution feature maps and enhance capturing the contextual information of the tiny lesions lost after multiple scales. We add the Atrous Spatial Pyramid Pooling (ASPP) [12] as a bridge between the ResNet encoder and the dual-decoder resolutions in the EDBNet framework.

2.3. Head: The Decoder Layer

The main new component of our methodology is the design of the head part. EDBNet employs a shared ResNet50 encoder to improve the performance without computation, and memory overload and dual-decoder boosted for lesions segmentation. The dual-decoder boosted has the same internal Gated Skip Connections (GSCs) architecture in each decoder block. Figure 2 presents the GSCs mechanism of each decoder block. GSCs modified the standard UNet decoder with a boosted feature maps production to enhance the discrimination between the lesion and background pixels for exudates segmentation.

EDBNet has four encoder blocks and four GSCs mechanism blocks for each decoder. In dec#1, the GSCs mechanism receives feature maps from the corresponding ResNet encoder layers. It concatenates them with the feature maps produced by the previous block (either the ASPP neck block or a previous dec#1 block). For dec#2, it receives feature maps from the corresponding dec#1 layers. Then concatenate them with the previous block's feature maps (either the ASPP neck block or a previous dec#2 block). We can express these feature maps as $S_1 \in \mathbb{R}^{h \times w \times f}$, and $S_2 \in \mathbb{R}^{h/2 \times w/2 \times 2f}$. Then, S_2 feeds to UpSampled2D transposed convolution layer with a kernel size of 2×2 to produce feature maps \hat{S}_2 . \hat{S}_2 and S_1 should have the same width and height to perform the concatenation process as follows:

$$C = \varphi_{1 \times 1}([S_1 || \hat{S}_2]) \quad (1)$$

In this expression, $\varphi_{1 \times 1}$ stands for the convolution operation with a kernel size of 1×1 and $||$ refers to the concatenation operation. C feature maps are fed to a *sigmoid* activation function to generate the weights ϑ , which helps to improve the discrimination between the lesion pixels and background pixels for EX or SE segmentation tasks. These weights are multiplied by the summation of S_1 , \hat{S}_2 as follows:

$$D = \vartheta * (S_1 + \hat{S}_2) \quad (2)$$

After that, the improved feature maps of D are fed into two convolution layers, batch normalization and rectified linear unit activation function. The final output blocks of both dec#1 and dec#2 are followed by a sigmoid activation function and fed into a fusion process by the average weighted aggregation to take the benefits of multiple information sources and generate an optimal joint lesion segmentation [13]. In addition, the fusion aims to produce one final output with a fewer output channels as follows:

$$M = (M_1 + M_2)/2.0 \quad (3)$$

Where M stands for the final output of the proposed framework, whereas M_1 and M_2 indicate the output masks of dec#1 and dec#2, respectively. The final output mask is a binary image that includes the EX or SE lesions and has a size identical to the size of the input image size (384×256).

3. Experimental Results and Discussions

This section describes the conducted experiments to evaluate the effectiveness of the proposed model, including the description of the dataset, the experimental setup, the evaluation metrics and the analysis of the obtained results.

3.1. Dataset, pre-processing, and experimental setup

We used the popular Indian Diabetic Retinopathy Image Dataset (IDRiD) in our experiments [14]. It composed of 81 high-resolution retinal fundus images of 4288×2848 . Each image contains at least one mask labelled as one out of four types of DR lesions EX, SE, MA, and HE. The dataset was split into 54 images as the training set and the rest of 27 as the testing set.

We used the following training pipeline (including some data augmentation techniques to enrich the data and improve the regularity of the model) to process the images in the training set. First, each image is divided into four non-overlapped sub-images, and the corresponding sub masks are constructed. We ignored the negative sub-images, i.e., the sub-images only with the background mask. Hence, each example in the training process is a sub-image with the size of 2144×1424 pixels with its corresponding sub mask. Next, we resized the sub-images and the sub masks to 384×256 . The interpolation mechanism used is cubic for the images and the nearest neighbour for the masks.

After that, we applied horizontal flipping, rotation, Gaussian noise and grayscale augmentation techniques for 12 times. The total of training data calculated as: $((4 \times 54) - 20\%) \text{ (for validation)} \times 12$.

We trained each model for 50 epochs using Adam optimizer and a batch size of 4. The learning rate is set to 0.001. We sampled a subset (20%) from the training set and

used it as a validation set to save the best checkpoint of the trained models. We used the binary cross-entropy as a loss function to train the models.

During the inference phase, we only resize the input image to 768×512 and perform a full image segmentation process (i.e., no image splitting or image augmentation is used during the inference).

3.2. Evaluation Metrics

In this study, we used the most common evaluation metrics in the lesions segmentation task, specifically:

- Area Under Precision-Recall curve (AUPR): it is known to be a realistic measure for lesion segmentation performance like exudates [15].
- Pixel Accuracy (ACC): It can be defined as the percent of pixels in the image which were correctly classified. Formally it is defined as the following:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

- Intersection-over-Union (IoU): also known as the Jaccard index, is basically a method to compute the percent overlap between the ground truth mask and the predicted mask.

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (5)$$

- Dice Coefficient: also referred to as F-score, it is the harmonic mean of precision and recall. It can be expressed as follows:

$$\text{Dice} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \quad (6)$$

- TPR/Sensitivity: denotes the proportion of real lesion pixels classified as lesion pixels. Formally it is defined as the following:

$$\text{TPR/Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

- FNR: denotes the model incorrectly predicts the negative class (background pixels). Formally it is defined as the following:

$$\text{FNR} = 1 - \text{TPR} \quad (8)$$

In these expressions, TP refers to the true positive (the pixels were labelled as foreground, i.e., retinal lesion pixels, and correctly classified). The term FP means false positive (the pixels were labelled as background and misclassified as foreground). Also, TN is a true negative, which refers to the healthy pixels correctly classified by the network. whereas FN is a false negative representing the lesion pixels misclassified as healthy pixels. Finally, the TPR refers to the true positive rate of TP lesion pixels prediction, and the FNR indicates the false negative rate.

3.3. Different Decoder Architectures Evaluation of Retinal Exudates Segmentation

In this section, we evaluate the performance of the used decoder architectures on the test set images of the IDRiD dataset. Table 1 presents the performance of EX and SE retinal lesion segmentation models with the IDRiD dataset. We conducted three different experiments of decoder-side architecture for each EX and SE retinal lesions segmentation—Unet, UNet + GSCs, and EDBNet. We achieved the result of 74.4% of

Table 1. Performance comparison on IDRiD dataset. Values in bold highlight indicates to the highest case of accuracy

Methods	SE				EX			
	ACC	AUPR	IOU	Dice	ACC	AUPR	IOU	Dice
UNet	99.8±0.001	70.7±0.289	74.3±0.064	85.3±0.049	99.5±0.002	81.4±0.149	76.5±0.024	86.7±0.015
UNet + GSCs	99.8±0.001	73.4±0.271	76.5±0.071	86.7±0.053	99.5±0.002	82.0±0.149	78.1±0.023	87.7±0.014
EDBNet	99.8±0.001	74.4±0.273	78.0±0.072	87.6±0.053	99.5±0.002	85.3±0.157	80.3±0.009	89.1±0.006

SE lesion segmentation and 85.3% of EX lesion segmentation with the AUPR metric that common uses of the IDRiD dataset challenge. For the Accuracy metric of SE and EX, we reached 99.8 and 99.5%, respectively. Regarding the Dice metric, we obtained 87.6% of SE and 89.1% of EX. We noted our GSCs mechanism enhanced the performance of SE and EX lesions segmentation when it has added to the UNet decoder, as shown in the second row of Table 1. At the same time, when we used the EDBNet, the performance of SE and EX lesions segmentation improved compared to the method that just used GSCs in one decoder, as shown in the third row. As a result, the SE lesion segmentation performance has improved by 1%, 1.5%, and 0.9% of AUPR, IOU, and Dice metrics, respectively.

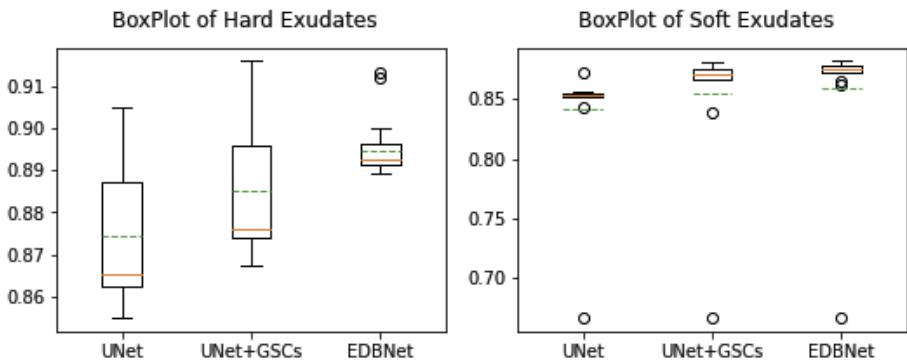


Figure 3. BoxPlot of Dice for Hard and Soft Exudates Segmentation results (green dashed lines indicate the mean, and the oranges indicate the median). All values outside the whiskers are considered as outliers, which are marked with the (o) symbol.

On the other hand, the EX lesion segmentation performance has significantly enhanced the AUPR, IOU, and Dice metrics of 3.3%, 1.8%, and 1.4%, respectively.

The EDBNet reduced the standard deviation from ± 0.015 to ± 0.006 with Dice and ± 0.024 to ± 0.009 with IOU metrics of EX segmentation. These effects reveal that EDBNet can present more precise and robust segmentation than the UNet and UNet + GSCs models.

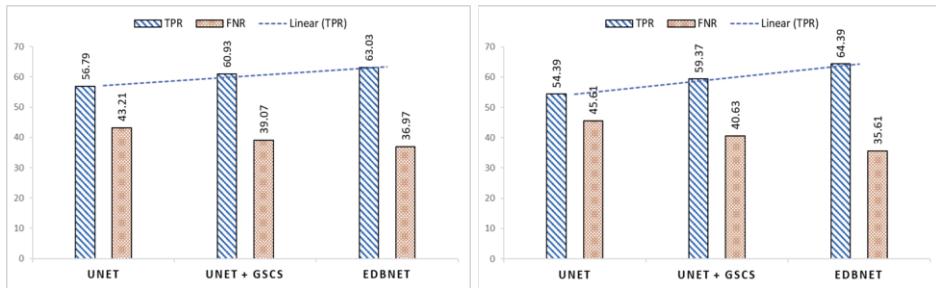


Figure 4. Average of TPR and FNR with different decoder architectures (EX left, SE right).

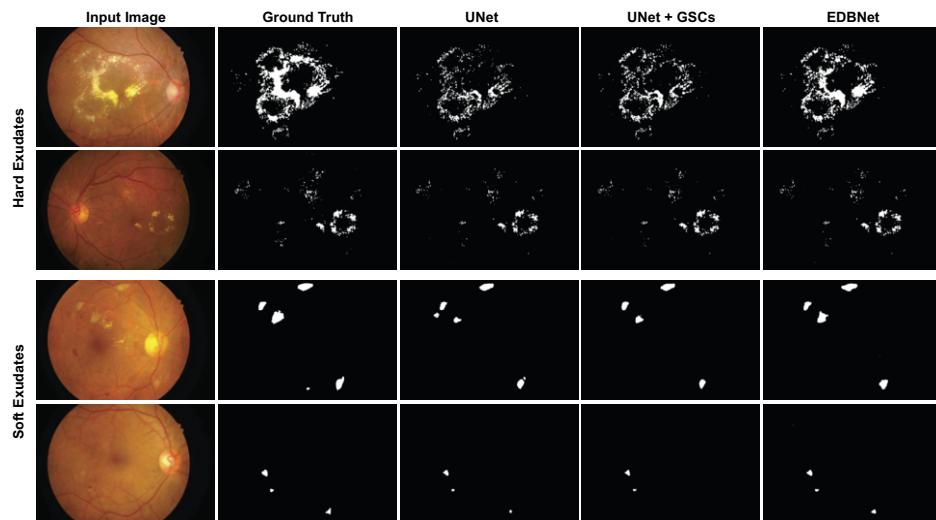


Figure 5. Hard and Soft Exudates Segmentation results.

Figure 4 shows that the EDBNet model of the dual-decoder yields the highest True Positive Rate (TPR/Sensitivity) and lowest True Negative Rate (TNR) for all testing set images of the IDRiD dataset. The dashed line indicates the performance enhancement when we used the GSCs decoder and the EDBNet. In addition, we present some sample masks obtained for the three SE and EX lesion segmentation models to demonstrate the efficacy as shown in Figure 5. Figure 5 shows that the final output masks of EDBNet are better segmented than those final output masks of UNet+GSCs and UNet baseline models.

Finally, we show the boxplots of the Dice of the proposed model, UNet, and UNet+GSCs. As shown in Figure 3 among the tested models, the proposed model has the highest mean, median, and smallest standard deviation of the Dice for the EX and has the highest mean and median for the SE.

It is engaging to determine the statistical significance of the differences in performance between the proposed EDBNet and UNet + GSCs (the second best model) in terms of Dice. To do so, we used Student's t-test (significance level < 0.05) to specify

the difference between Dice values. The p-values obtained are less than 0.05 for EX and higher than 0.05 for SE, indicating a statistical significance for Ex but not for SE.

3.4. Comparison with existing methods

To confirm the effectiveness of the proposed method, we conduct a comparison between EDBNet and state-of-the-art on the AUPR metric (same as the one used in the IDRiD competition) as shown in Table 2. The comparison includes the top 5 teams on the IDRiD competition [16] (the first five rows in the table), as well as CARNet [4], EAD-Net [5], L-Seg [17], and SAA [7]. EDBNet surpasses all the state-of-the-art results of segmenting the SE retinal lesion segmentation by 1.55% of the AUPR metric. Regarding the EX, EDBNet achieves acceptable performance. The results of EDBNet

Table 2. Comparison with state-of-the-art results on IDRiD

Method	AUPR on SE	AUPR on EX
VRT (1st) [16]	0.6995	0.7127
PATech (2nd) [16]	-	0.8850
IFLYTEK-MIG (3rd) [16]	0.6588	0.8741
SOONER (4th) [16]	0.5395	0.7390
SAIHST (5th) [16]	-	0.8582
CARNet [4]	0.7125	0.8675
EAD-Net [5]	0.6083	0.7818
L-Seg [17]	0.7113	0.7945
SAA [7]	0.7281	0.8792
Proposed	0.7436	0.8534

are comparable with IFLYTEK-MIG, CARNet, and SAA, while they had good results of EX lesion segmentation but not with SE. On the other hand, PATech and SAIHST have the best results for EX retinal lesion segmentation but did not introduce SE retinal lesion segmentation results to compare. Therefore, as shown in Table 2 there is no method with the best results for the two retinal lesions segmentation at the same time.

4. Conclusions and future work

This paper presented a new deep learning architecture called EDBNet for image segmentation problems. It has been used for segmenting retinal exudates in fundus images of the human eye. EDBNet is composed of three main elements: ResNet50 backbone, ASPP neck, and dual-decoder using several GSCs in cascade. EDBNet framework led to high segmentation performance of both SE and EX retinal lesions for the IDRiD dataset, obtaining an accuracy, AUPR, IOU, and Dice metrics of 99.8, 74.4, 78.0, and 87.6% of soft exudates, respectively; while for hard exudates, we achieve 99.5, 85.3, 80.3, and 89.1%, respectively. EDBNet showed superiority with respect to other state-of-the-art models, with performance higher than 1.55% of the AUPR metric for the SE retinal lesion segmentation with the IDRiD dataset.

The EDBNet framework is not designed exclusively for retinal lesion segmentation on fundus images; it may also work well in cases of medical images with tiny objects.

Future work will include using the proposed exudates segmentation model to develop segmentation methods for four kinds of eye lesions: microaneurysm (MA), haemorrhage (HE), hard exudate (EX), and soft exudate (SE) to diagnose diabetic retinopathy.

Acknowledgements

This work has been funded by the research projects PI21/00064 and PI18/00169 from Instituto de Salud Carlos III & FEDER funds. The University Rovira i Virgili also supports this work with projects 2020PFR-B2-61 and 2019PFR-B2-61.

References

- [1] Mary VS, Rajsingh EB, Naik GR. Retinal fundus image analysis for diagnosis of glaucoma: a comprehensive survey. *IEEE Access*. 2016.
- [2] Jani K, Srivastava R, Srivastava S, Anand A. Computer aided medical image analysis for capsule endoscopy using conventional machine learning and deep learning. In: 2019 7th International Conference on Smart Computing & Communications (ICS&C). IEEE; 2019. p. 1-5.
- [3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234-41.
- [4] Guo Y, Peng Y. CARNet: Cascade attentive RefineNet for multi-lesion segmentation of diabetic retinopathy images. *Complex & Intelligent Systems*. 2022;1:1-21.
- [5] Wan C, Chen Y, Li H, Zheng B, Chen N, Yang W, et al. EAD-net: a novel lesion segmentation method in diabetic retinopathy using neural networks. *Disease Markers*. 2021;2021.
- [6] Li Y, Zhu M, Sun G, Chen J, Zhu X, Yang J. Weakly supervised training for eye fundus lesion segmentation in patients with diabetic retinopathy. *Mathematical Biosciences and Engineering*. 2022;19(5):5293-311.
- [7] Bo W, Li T, Liu X, Wang K. SAA: Scale-Aware Attention Block For Multi-Lesion Segmentation Of Fundus Images. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE; 2022. p. 1-5.
- [8] Jabreel M, Abdel-Nasser M. Promising crack segmentation method based on gated skip connection. *Electronics Letters*. 2020;56(10):493-5.
- [9] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44.
- [10] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.
- [11] Pan X, Jin K, Cao J, Liu Z, Wu J, You K, et al. Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. *Graefes Archive for Clinical and Experimental Ophthalmology*. 2020;258(4):779-85.
- [12] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*. 2017;40(4):834-48.
- [13] Zhang J. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*. 2010;1(1):5-24.
- [14] Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*. 2018;3(3):25.
- [15] Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2013. p. 451-66.
- [16] Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. Idrid: Diabetic retinopathy-segmentation and grading challenge. *Medical image analysis*. 2020;59:101561.
- [17] Guo S, Li T, Kang H, Li N, Zhang Y, Wang K. L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images. *Neurocomputing*. 2019;349:52-63.

This page intentionally left blank

Explainability and Argumentation

This page intentionally left blank

Combining Support and Attack Interactions for Argumentation Based Discussion Analysis

Teresa ALSINET¹, Josep ARGELICH and Ramón BÉJAR

*INSPIRES Research Center – University of Lleida
Jaume II, 69 – 25001 Lleida, SPAIN*

Abstract

Online discussion is one of the commonly used tools to engage users to discuss relevant topics for society, where user contributions in the form of posts, comments and votes are essential to their success. The scale, complexity and dynamism of this information leads to a growing interest in understanding what are the major accepted or rejected opinions in different domains by social network users. In this work, we explore how to combine attack and support interactions to extract possible consensus based on some abstract argumentative semantics, and we characterize the set of properties or postulates that this consensus should satisfy.

Keywords. Online discussion, probabilistic weighted interactions, valued arguments, support and attack relations, postulates for consensus analysis.

1. Discussion model

Our goal in this work is to consider a general online discussion platform and to reason, by means of an argumentative approach, about the set of posts that can be accepted as consensus among the participants by combining both the social relevance of posts and the degrees of belief in the answers between them.

Argumentation includes various forms of dialogue such as deliberation and negotiation, which are concerned with collaborative decision-making procedures by which people can express and rationally resolve or at least manage their disagreements. An abstract argument framework, as proposed by Dung [11], is a graph structure in which the nodes denote arguments and the edges denote attacks between the arguments. When we are considering online discussions, another kind of interaction may exist between posts. Indeed, a post can attack another post, but it can also support another one. In addition, it is common for each post to have a degree of popularity, preference or social support, such as the votes it receives throughout the discussion.

To represent the characteristics of online discussions, our proposal is based on graphs extended with weights for both nodes and edges. Each post gives rise to a node in the

¹Correspondence to: T. Alsinet. INSPIRES Research Centre, University of Lleida. C/Jaume II, 69. Lleida, Spain. Tel.: +34 973702734; E-mail: teresa.alsinet@udl.cat.

graph, and relationships between posts give rise to the edges of the graph expressing answers between them. We say that a post p_1 *answers* a post p_2 whenever p_1 is a direct reply to p_2 or p_1 mentions (refers to) p_2 . So, a post can answer many posts. In what follows, Γ denotes a non-empty set of posts, and it is referred to as a discussion from a root (main) post.

Definition 1 (*Weighted Discussion Graph*) A weighted discussion graph (*WDisG*) for a discussion Γ is a directed graph $G = \langle N, E, L, W \rangle$, where

- For every post p_i in Γ , there is a node n_i in N .
- If post p_1 answers post p_2 , there is a directed edge (n_1, n_2) in E .
- L is a labelling function $L : E \rightarrow [0, 1]^3$ for edges in E . The labelling function L maps an edge (n_1, n_2) to a triple of probability values $(p_a, p_s, p_n) \in [0, 1]^3$ with $p_a + p_s + p_n = 1$, which expresses the probability or degree of belief that the answer from post p_1 to post p_2 can be classified as attack (p_a), support (p_s) and none (p_n), respectively. Attack means that the comment expressed in the post p_1 criticizes or disagrees with the claim expressed in the post p_2 , support that p_1 agrees with the claim expressed in p_2 and none that the relation is none of the previous two.
- W is a weighting function $W : N \rightarrow [0, 1]$ for nodes in N . The weighting function W maps the node n_i of a post p_i in Γ to a value in the real interval $[0, 1]$, which expresses the social (support) acceptance degree that the comment on the post has received throughout the discussion with respect to the rest of comments.

2. Support and attack interactions

Once we have introduced the formal representation of discussions as Weighted Discussion Graphs, the next key component is the classification of the interactions between nodes according to the labelling function L of edges and the weighting function W of nodes. In our approach, we define two types of interactions between nodes, which we refer to as attack (R_{att}) and support (R_{sup}) relations, and we consider an uncertainty threshold $\alpha \in (0, 1]$ which characterizes how much uncertainty about classification we are prepared to tolerate.

Fuzzy argumentation frameworks deal with the uncertainty of arguments or attacks caused by incompleteness or vagueness, and its semantics have been studied in various ways. In [12] the authors propose a fuzzy argumentation approach by enriching the expressive power of the classical argumentation model proposed by Dung [11], by allowing to represent the relative strength of the attack relationships between arguments, as well as the degree to which arguments are accepted. They define extensions as fuzzy sets by aggregating the strength of the attack with the degree of acceptance of arguments. Although our approach is not based on fuzzy sets semantics, we also propose to use an aggregation operation to define the attack and support relations.

Definition 2 (*Support and Attack Interactions*) Let $G = \langle N, E, L, W \rangle$ be a *WDisG* for a discussion Γ and let $\alpha \in (0, 1]$ be a threshold on the distances of the probability values. Moreover, let $\wedge : [0, 1] \times [0, 1] \rightarrow [0, 1]$ be a t-norm; i.e., a binary aggregation operation satisfying: monotone, associative, commutative, and $\forall x \in [0, 1], 1 \wedge x = x \wedge 1 = x$. We define two binary relations on N as follows:

- A binary relation R_{sup} on N called the support relation:

$$R_{sup} = \{(n_1, n_2) \in E \mid L(n_1, n_2) = (p_a, p_s, p_n) \text{ with } p_s \geq \max(p_a, p_n) + \alpha \text{ and } W(n_1) \wedge p_s \geq W(n_2)\}.$$
- A binary relation R_{att} on N called the attack relation:

$$R_{att} = \{(n_1, n_2) \in E \mid L(n_1, n_2) = (p_a, p_s, p_n) \text{ with } p_a \geq \max(p_s, p_n) + \alpha \text{ and } W(n_1) \wedge p_a \geq W(n_2)\}.$$

Notice that the support and attack relations verify that $R_{sup} \cap R_{att} = \emptyset$ and $R_{sup} \cup R_{att} \subseteq E$; i.e., the answer between two posts is neither a support nor an attack interaction, or, if it is, it is either a support or an attack interaction, but not both.

Finally, from R_{sup} and R_{att} , we recursively define the set of posts that reinforce (*support*) or contradict (*attack*) a post $n_i \in N$ in the following way:

- $support(n_i) = \{n_i\} \cup \{n_1 \mid (n_1, n_2) \in R_{sup} \cap (N \times support(n_i))\}$
- $attack(n_i) = \{n_1 \mid (n_1, n_2) \in R_{att} \cap (N \times support(n_i))\}$

Then, we say that a discussion Γ is *coherent* whenever $support(n_i) \cap attack(n_i) = \emptyset$, for all node $n_i \in N$.

3. Rationality postulates

The basic idea of argumentation theory is to construct arguments in favour and against a statement (in our approach, a post), to select the “acceptable” ones, accepted as consensus among the participants, and, finally, to determine whether the original statement (root post) can be accepted or not.

In [1] we defined a reasoning system for analysis of discussions on Twitter, where each tweet is represented by an argument and the notion of acceptability is based on Value-based Abstract Argumentation. Value-based Abstract Argumentation [6], attach arguments with social values, and makes the semantics dependent on a particular preference order over values, representing a particular audience. While our reasoning system is useful in applications as user profile analysis [2], it can lead to some unintuitive results, since the acceptability semantics does not consider support relationships between arguments. In [1] the authors take into account support interactions by reinforcing the social values of arguments propagating support relationships between tweets.

In order to consider a more intuitive acceptability semantics for online discussions, in this preliminary work, we are interested in defining principles, as the rational postulates proposed by Caminada and Amgoud [8] for rule-based argumentation systems, [7] for non-monotonic reasoning with strict and defeasible rules, and [4] for weighted bipolar settings.

Definition 3 (Postulates) Let R_{sup} and R_{att} be the support and attack relations, respectively, for a WDisG $G = \langle N, E, L, W \rangle$ of a coherent discussion Γ . An acceptable consensus $A \subseteq N$ of Γ should satisfy:

- *Consistency*, meaning that $\neg \exists (n_1, n_2) \in A \times A : \exists n_3 \in attack(n_2) \text{ and } n_1 \in support(n_3)$.
- *Closure*, meaning that $\neg \exists n_i \in A : support(n_i) \not\subseteq A$.
- *Self-defence*, meaning that $\forall (n_1, n_2) \in R_{att} \cap ((N \setminus A) \times A) : \exists (n_3, n_4) \in R_{att} \cap (A \times support(n_1))$.

Notice that *consistency* ensures both *direct* (i.e., $\neg\exists (n_1, n_2) \in A \times A : n_1 \in \text{attack}(n_2)$) and *indirect* consistency (i.e., $\neg\exists (n_1, n_2) \in A \times A : \exists n_3 \in \text{attack}(n_2) \text{ and } n_1 \in \text{support}(n_3)$).

Finally, another feature or postulate that should satisfy an acceptable consensus $A \subseteq N$ is that of maximality, meaning that $\neg\exists n_i \in (N \setminus A) : A \cup \{n_i\}$ satisfies consistency, closure and self-defence.

In the literature, we find different approaches to incorporate support between arguments in the context of Abstract Argumentation Frameworks. In [9], the authors introduce Bipolar Abstract Argumentation, extending the defeat relation with indirect attacks, and in [10] acceptability semantics are proposed based on enforcing coherence of the admissible sets and taking into account attack and support interactions for proposing gradual labelling for the arguments. On the other hand, weighted semantics [3,5] focus on the evaluation of individual arguments rather than sets of arguments and assign a weight to each argument, allowing for a fine-grained classification of the acceptance and rejection of arguments. As future work, we intend to explore these approaches in order to extract an order of preference for the possible acceptable consensus sets of a discussion.

Acknowledgments The authors would like to thank the anonymous reviewers for providing helpful comments. This work was partially funded by Spanish Project PID2019-111544GB-C22, by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 723596 and Grant Agreement 768824, and by 2017 SGR 1537.

References

- [1] Alsinet, T., Argelich, J., Béjar, R., Fernández, C., Mateu, C., Planes, J.: An argumentative approach for discovering relevant opinions in twitter with probabilistic valued relationships. *Pattern Recognition Letters* 105, 191–199 (2018)
- [2] Alsinet, T., Argelich, J., Béjar, R., Martínez, S.: User profile analysis in reddit debates. In: *Artificial Intelligence Research and Development*, CCIA 2019. vol. 319, pp. 275–284. IOS Press (2019)
- [3] Amgoud, L., Ben-Naim, J.: Ranking-based semantics for argumentation frameworks. In: *SUM 2013. Lecture Notes in Computer Science*, vol. 8078, pp. 134–147. Springer (2013)
- [4] Amgoud, L., Ben-Naim, J.: Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.* 99, 39–55 (2018)
- [5] Amgoud, L., Ben-Naim, J., Doder, D., Vesic, S.: Acceptability semantics for weighted argumentation frameworks. In: *IJCAI 2017*. pp. 56–62. ijcai.org (2017)
- [6] Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13(3), 429–448 (2003)
- [7] Caminada, M.: Rationality postulates: Applying argumentation theory for non-monotonic reasoning. *FLAP* 4(8) (2017)
- [8] Caminada, M., Amgoud, L.: On the evaluation of argumentation formalisms. *Artif. Intell.* 171(5-6), 286–310 (2007)
- [9] Cayrol, C., Lagasquie-Schiex, M.: On the acceptability of arguments in bipolar argumentation frameworks. In: *Proceedings of ECSQARU 2005*. pp. 378–389 (2005)
- [10] Cayrol, C., Lagasquie-Schiex, M.: Bipolar abstract argumentation systems. In: *Argumentation in Artificial Intelligence*, pp. 65–84. Springer (2009)
- [11] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2), 321 – 357 (1995)
- [12] Janssen, J., Cock, M.D., Vermeir, D.: Fuzzy argumentation frameworks. In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU'2008 (2008)

Focus and Bias: Will It Blend?

Anna ARIAS-DUART^{a,1}, Ferran PARÉS^a and Víctor GIMÉNEZ-ÁBALOS^a

Dario GARCIA-GASULLA^a

^aBarcelona Supercomputing Center (BSC)

Abstract.

One direct application of explainable AI feature attribution methods is to be used for detecting unwanted biases. To do so, domain experts typically have to review explained inputs, checking for the presence of unwanted biases learnt by the model. However, the huge amount of samples the domain experts must review makes this task more challenging as the size of the dataset grows. In an ideal case, domain experts should be provided only with a small number of selected samples containing potential biases. The recently published Focus score seems a promising tool for the selection of samples containing potential unwanted biases. In this work, we conduct a first study in this direction, analyzing the behavior of the Focus score when applied to a biased model. First, we verified that Focus is indeed sensitive to an induced bias. This is assessed by forcing a spurious correlation, training a model using only cats-indoor and dogs-outdoor. We empirically prove that the model learnt to distinguish the contexts (outdoor vs indoor) instead of cat vs dog classes, so ensuring that the model learnt an unwanted bias. Afterwards, we apply the Focus on this biased model showing how the Focus score decreases when the input contains the aforementioned bias. This analysis sheds light on the Focus behavior when applied to a biased model, highlighting its strengths for its use for bias detection.

Keywords. Focus, explainability, feature attribution methods, evaluation metrics, bias, mosaics

1. Introduction

Neural networks have proven to be effective tools for image classification tasks [1]. However, their interpretation is still obscure. The most popular methods used to overcome this problem are post-hoc attribution methods, such as SmoothGrad [2], GradCAM [3], Layer-Wise Relevance Propagation (LRP) [4] or LIME [5]. These methods provide an attribution map representing the contribution of pixels towards the final prediction. In order to assess the reliability of these techniques, different approaches have been proposed in the literature. Evaluation metrics such as the Pointing Game [6] or the Region Perturbation [7]. Recent works exploit the use of grids to carry out these evaluations [8,9].

In this work, we analyze in depth the behavior of the Focus [8] score, particularly when applied to a biased model. To do so, we trained a model on biased data, enabling it to learn a spurious correlation we can quantify and control. Then we analyse the Focus behaviour when applied to this biased model. Notice that these spurious correlations are difficult to detect and validate using classic performance metrics (e.g., loss, accuracy), since these unwanted biases helps the model to learn and obtain correct predictions.

¹Corresponding Author: Anna Arias-Duart; E-mail: anna.ariasduart@bsc.es.

2. Related Work

One of the powerful uses of the feature attribution methods is the detection of unwanted biases in datasets and models. However, deciding whether these spurious correlations are desirable or undesirable is for domain experts to say, following ethical and practical considerations. To enable experts to do this job, exploiting these attribution maps, we need to provide them with useful and reliable information.

Some work has been done in this direction, Lapuschkin *et al.* [10] propose to reduce the explanations space provided to the domain experts through spectral clustering, so producing a reduce set of clusters instead of thousands of explanations. Where these clusters aim to represent different classification strategies and then, these strategies are shown to domain experts for the final unwanted bias detection stage. Similarly, this work [8] proposes a methodology where Focus is used to reduce the number of explanations to a subset that highlights potential unwanted biases, hence considerably reducing the search space to be assessed by domain experts for finding unwanted biases. However, the authors in [8] use models in which they have no control over the actual bias, limiting the reliability of the results. In our work, we go beyond, applying the Focus on a model to which we induce an unwanted bias.

The Focus metric involves three elements: a feature attribution method, a trained classification model and a set of mosaic samples. Each mosaic is composed of images of different classes, some of them from a *target class*, the specific class the explainability method is expected to explain. For example, if the *target class* of a given mosaic is the *dog* class, at least one of the images within the grid of the mosaic must belong to the *dog* class (see Figure 3 for examples of 2×1 mosaics). Intuitively, if we ask a feature attribution method for the explainability of the *target class* on a mosaic, the Focus metric will measure the proportion of attribution lying on the *target class* squares, with respect to the total mosaic attribution. A random (uniform) attribution obtains a Focus equal to the proportion of squares.

In this context, we analyze the Focus behavior when applied to a model where the spurious correlation learnt by the model is known, verifying its potential use as a tool for bias detection.

3. Building a biased model

To apply the Focus on a biased model, we first need a biased dataset to learn from. In this section, we first explain how we created the biased dataset §3.1. Then, we introduce the training configurations §3.2. And last but not least, we perform some sanity checks to confirm that indeed we managed to introduce a spurious correlation into the model §3.3.



Figure 1. Examples of indoor/outdoor images: (a) cat-indoor (b) cat-outdoor (c) dog-indoor (d) dog-outdoor.

Table 1. Contexts included in each category for the Visual Genome dataset. The first and second column corresponds to the cat-outdoor and cat-indoor category. And the third and forth column to the dog-outdoor and dog-indoor category respectively.

				
car, fence, grass, roof, bench, bird, house	speaker, computer, screen, laptop, computer mouse, keyboard, monitor, desk, sheet, bed, blanket, remote control, comforter, pillow, couch, books, book, television, bookshelf, blinds, sink, bottle, faucet, towel, counter, curtain, toilet, pot, carpet, toy, floor, plate, rug, food, table, box, paper, suitcase, bag, container, vase, shelf, bowl, picture, papers, lamp, cup, sofa	house, grass, horse, fence, cow, sheep, dirt, car, motorcycle, truck, helmet, snow, flag, boat, rope, trees, frisbee, bike, bicycle, sand, surfboard, water, fire hydrant, pole, skateboard, bench, trash can	screen, shelf, desk, picture, laptop, remote control, blanket, bed, sheet, lamp, books, pillow, curtain, container, table, cup, plate, food, box, rug, floor, cabinet, towel, bowl, television, carpet, sofa	

3.1. Dataset creation

The creation of this dataset is motivated by the need to have control over some of the dataset biases. To do so, we use the MetaShift [11] to induce a correlation that we can quantify and control. This work clusters the images according to metadata. An annotated graph is created where each node represents a class in a specific context, for example *dog frisbee*. The distance between nodes represents the similarity between those contexts: *dog frisbee* will be closer to *dog grass* than *dog books*. The more contexts are shared within a class, the closer the nodes will be. Using the construction proposed by [11] we create a dataset composed of two classes (cat and dog) with two subclasses (indoor and outdoor), see Figure 1 for details.

We built the dataset with images from two well-known datasets, both providing contextual information: the Common Objects in Context (COCO) dataset [12] and the Visual Genome dataset [13]. Tables 1 and 2 show the exact contexts used for the construction of the indoor and outdoor subclasses for both datasets, the Visual Genome dataset and the COCO dataset respectively.

Table 2. Contexts included in each category for the COCO dataset. The first column corresponds to the outdoor contexts and the second to the indoor ones.

		
bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket	bottle, wine glass, cup, fork, knife, spoon, bowl, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush	

3.2. Model

Next, we train a model using only samples from cats-indoor and dogs-outdoor. In this way, we introduce a spurious correlation, which could in fact occur in a real scenario: dog-outdoor images are more likely than cat-outdoor images.

For training the model, we use a total of 1,060 images per class (cats-indoor vs dogs-outdoor). Where 960 images per class were used for training and 100 for validation. We use the ResNet-18 [14] architecture, the AMSGrad [15] to optimize weights and we perform data augmentation during training: random rotation ($[-30, 30]$ degrees), random crop and random horizontal flip with a chance of 50%. We reach a mean per class accuracy on the validation set of 87%, corresponding to the model with the minimum validation loss. From here we will call this model: the *biased model*.

For comparison purposes, we also train a model avoiding the context correlation. We use the same training size (1,060 images per class) but in this case both cats and dogs will be equally present in both contexts 50% outdoors and 50% indoors. We reach a mean per class accuracy on the validation set of 60.5%, using the model with the minimum validation loss. Notice that the performance obtained is much lower, indicating that the induced context served as a successful shortcut to the model. Without this added bias, the high variability (different breeds) as well as the low quality (mislabeled samples or partially occluded animals) of the dataset, limits the performance of the model which fails to learn to distinguish the two classes robustly. From now on we will refer to this second model as the *non-biased model*. Both trainings are performed in a single computing node of the CTE-Power9 cluster at the Barcelona Supercomputing Center, with the following characteristics:

- $2 \times$ IBM Power9 8335-GTH @ 2.4GHz (20 cores and 4 threads/core).
- $4 \times$ GPU NVIDIA V100 (Volta) with 16GB HBM2.

3.3. Sanity checks

To prove that the previous model, trained for the cats (indoor) and dogs (outdoor) classification task, is biased indeed (i.e., it has managed to learn the context instead of cat and dog characteristic patterns), we perform the following experiment. Starting from the hypothesis that images predicted with low probability, or that are predicted as the opposite class (in the case of a binary classification problem) are likely to be those that have patterns of the opposite class, we selected the three dog images with the lowest prediction and the three worst cat image predictions, see Figure 2.

Before continuing with the hypothesis evaluation, it is worth mentioning how the samples predicted with least certainty significantly differ between cats and dogs. While for dogs, the lowest probability corresponds to 56.58% and the third lowest to 82.11% (both of which account for a correct classification in a binary problem), for cats these probabilities drop to 0.38% the lowest, and the third lowest to 47.29%. As shown in Figure 2 (and mentioned before) the worst predicted cat sample seems like a labeling mistake (labeled as cat indoor when it seems to be cat outdoor). We do not correct this mistake for the sake of methodological consistency. These results show a higher performance when classifying outdoor-dogs than indoor-cats, suggesting that the model has learnt to focus more on outdoor than indoor patterns. This may be due to the fact that outdoor patterns

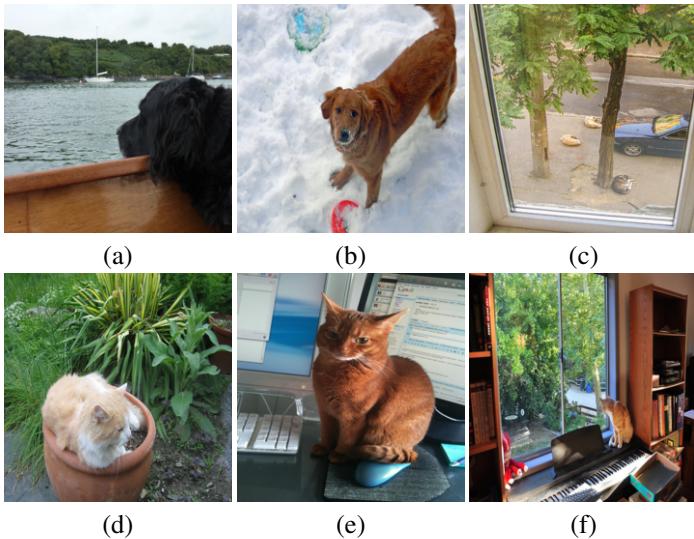


Figure 2. Examples of the worst predictions of the validation images set. Worst dog predictions: (a) dog: 0.5658, (b) dog: 0.7948 and (c) dog: 0.8211. Worst cat predictions: (d) cat: 0.0038, (e) cat: 0.4627 and (f) cat: 0.4729.

are less variant and more frequent, being a perfect visual pattern to discriminate between both classes.

Following the previous hypothesis: images predicted with a low probability are those that most likely contain a pattern of the opposite class. We build a set of 2×1 mosaics by combining those pairs of images (cats vs dogs), in order to apply a feature attribution method on top of them. Notice that the use of feature attribution methods on top of the mosaics enhances the detection of shared biases (term introduced in [8]). The shared biases are characteristic patterns of one class that are present in another class.

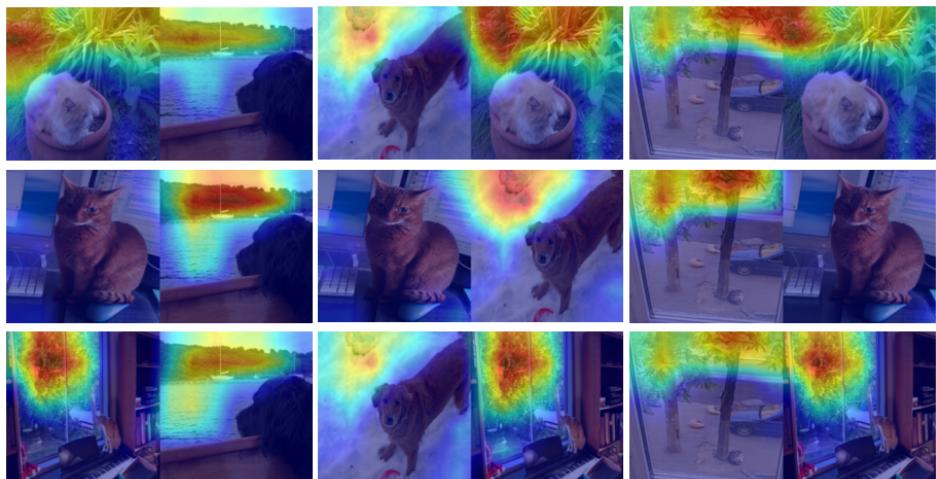


Figure 3. Feature attribution maps obtained by GradCAM on the *bias model* (the dog being the *target class*).

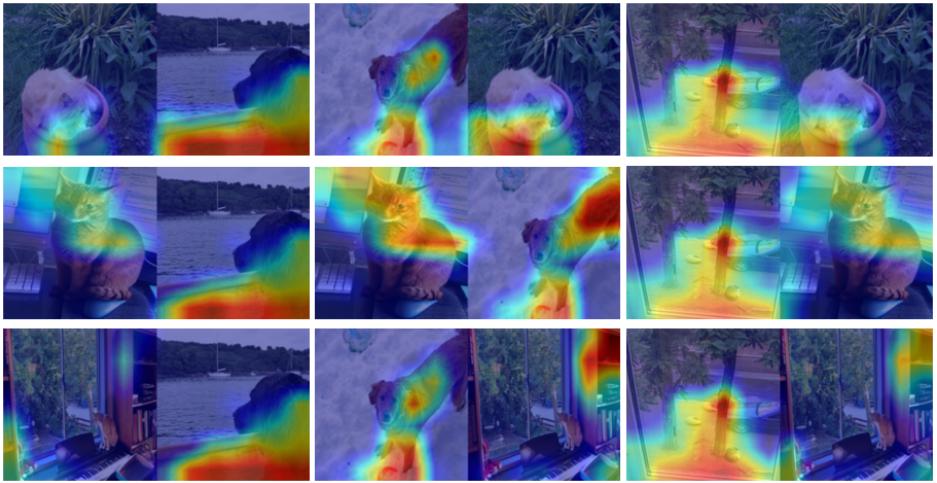


Figure 4. Feature attribution maps obtained by GradCAM on the *bias model* (the cat being the *target class*).

For this experiment we use the GradCAM attribution method. Results for both *target classes* are shown in Figure 3 and Figure 4. In all mosaics with the *target class* being the dog (see Figure 3), the GradCAM attribution focuses on areas where trees, leaves, or plants are present. Regardless of these patterns appearing in the cats squares or in the dog squares. Based on this, we hypothesize that the model has learnt to detect vegetation, instead of discriminating between cats and dogs. Indeed, it seems reasonable to think that most dogs in an outdoor context will be on meadows, fields or mountains (with a prominent presence of vegetation), while indoor cats will lack such pattern. This situation would have made it easier for the model to distinguish between dogs and cats, by only learning the green context instead of learning the characteristic patterns of these two mammals.

Similarly, the attribution in Figure 4, with the cat being the *target class*, falls on the wood or the brown areas (see for example first column of Figure 4). This pattern, although to a lesser extent than the vegetation, seems to be learnt by the model as a characteristic pattern of the cat class.

In order to corroborate that the model has learnt to identify vegetation as a characteristic pattern of the dog class and the wood as characteristic of the cat class, we perform another sanity check. We fed the model with the hand selected images shown in Figure 5, obtained from external sources. Image (a) is an image of only grass, which is predicted as a dog with a probability of 99.98%. On the contrary, Image (b) is a wood image which is predicted as a cat with a probability of 96.30%. In the case of Image (c), both patterns are present, although the green pattern is more prominent. This image is predicted as a dog with a probability of 99.44%. Notice how the attribution, being the *target class* the dog class (see Image (g)), falls on the green part around the path. However, when we ask for the attribution of the cat class (see Image (h)), the relevance focuses on the wooden bridge.

These results validate our hypothesis: vegetation is the main pattern learnt by the model as characteristic of the dog class and the wood pattern is learnt as characteristic of the cat class. At this point, we can confirm that the model is clearly skewed, it has learnt

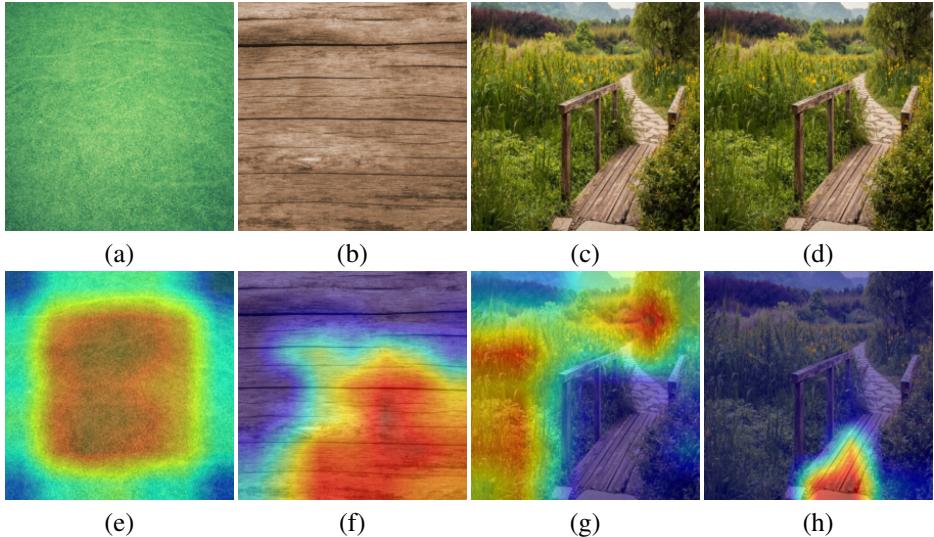


Figure 5. First row: hand selected samples predicted by the *biased model* as (a) dog: 0.9998 (b) cat: 0.9630 (c) and (d) dog: 0.9944. Second row: feature attribution maps obtained by GradCAM for the images of the first row being the *target class*: (e) the dog class (f) the cat class (g) the dog class and (h) the cat class.

to differentiate the two classes mainly by context and not by animal, and furthermore, we are aware of the principal patterns enabling such distinction.

4. Focus on a biased model

To evaluate the performance of the Focus score on a biased model we need three elements: a feature attribution method, a trained model and a set of mosaics. As an explainability method we use GradCAM, the one obtaining the best results for Focus [8]. As a trained classification model we use the ones introduced in §3.2 (the *biased model* and the *non-biased model*). Finally, for the mosaics, we build four sets of 2×1 mosaics, following all possible combinations. Notice each set contains the same amount of mosaics (10,000):

1. **cat-indoor vs dog-outdoor:** Combines 100 cat-indoor images and 100 dog-outdoor images. Note that this set follows the same distribution used for training the *biased model*.
2. **cat-indoor vs dog-indoor:** Combines 100 cat-indoor images and 100 dog-indoor images.
3. **cat-outdoor vs dog-indoor:** Combines 100 cat-outdoor images and 100 dog-indoor images. Note that this set corresponds to a distribution complementary to the one used for training the *biased model*.
4. **cat-outdoor vs dog-outdoor:** Combines 100 cat-outdoor images and 100 dog-outdoor images.

Note that none of these sets corresponds to the distribution used for training the *non-biased model* in which samples of all sets are used (cats and dogs equally sampled from indoor and outdoor contexts). At this point we can now compute the Focus obtained by

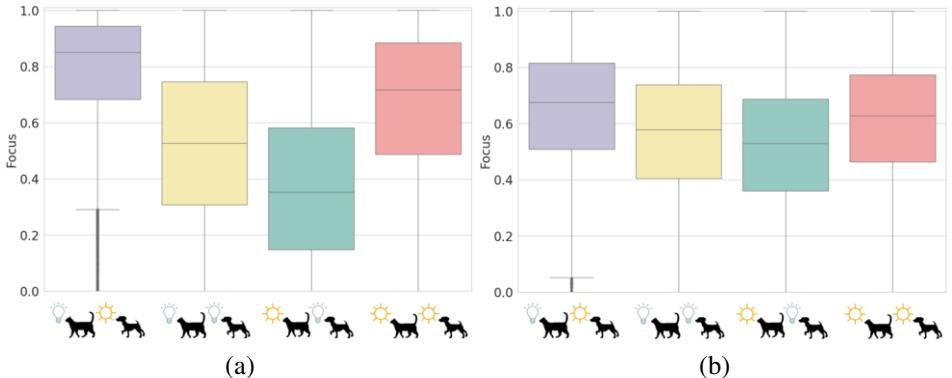


Figure 6. Each box plot shows the Focus distribution for a different validation set (evaluating 10,000 mosaics per set). The purple box plots correspond to the cat-indoor and dog-outdoor set (Set 1). The yellow box plots correspond to the cat-indoor and dog-indoor set (Set 2). The green box plots to the cat-outdoor and dog-indoor set (Set 3). And the red box plots to the cat-outdoor and dog-outdoor set (Set 4). (a) Focus distributions obtained by GradCAM on the *biased model* (b) Focus distributions obtained by GradCAM on the *non-biased model*.

each of the two models on each of the four mosaic sets. The resulting Focus distributions (including the 10,000 samples per set) are shown in Figure 6.

In the experiments with the *biased model*, the highest Focus is expected to be obtained with the Set 1, since the images within this set follow the same distribution in which the model has been trained. On the other hand, the Focus obtained with the Set 3, should be the lowest, since the images correspond to the completely inverse distribution. In this case, the mean Focus is expected to be between 0 and 0.5 since the learnt biases may be found on the the non *target class* squares.

In the experiments with the *non-biased model*, we expect the Focus distributions to be similar to one another. The training distribution of this model avoid biases regarding indoor and outdoor, which should prevent the model from focusing on these properties. Thus, if the context is not a factor, the four sets become analogous.

As seen in Figure 6, results follow our hypothesis. The context (indoor/outdoor) plays a significant role in the *biased model*, and have a much weaker impact on the results of the *non-biased model*. For the *biased model*, a mean Focus greater than 0.8 is obtained when using the same context as in training (Set 1, see first box plot in Figure 6 (a)). However, when the complementary distribution is used, Set 3, the mean Focus falls below 0.4. As hypothesized before, this low Focus is most likely due to the model finding patterns in the image of the opposite *target class*. Finally, the two sets having at least one correct context (Set 2 and Set 4) obtain a mean Focus in between the two mentioned above (see the second and the fourth box plot in Figure 6 (a)).

We hypothesize that a significant amount of label noise is found (particularly in the cat outdoor class, incorrectly labeling indoor cat images as outdoor samples). This would explain the fact that outdoor cats and dogs (red box plot of Figure 6 (a)) obtains a higher Focus than indoor cats and dogs (yellow box plot of Figure 6 (a)) as well as why the inverse distributed set (green box plot of Figure 6 (a), mean Focus of 0.3532) is not the complementary of the equally distributed set (purple box plot of Figure 6 (a), mean Focus of 0.8507).

In contrast, the Focus distributions obtained with the *non-biased model* have a mean Focus close to each other. The mean Focus obtained with Set 1 is still the highest, as

shown in Figure 6 (b), and the mean Focus obtained with Set 3 is slightly the lowest. This is likely to be caused by label noise induced by the natural predominance of cats to be indoor, and of dogs to be outdoor.

5. Conclusions

In this paper we analyze the behavior of Focus when applied to a biased model. To do so, we train a model to classify cats and dogs, to which we induce a correlation: we only use cats-indoor and dogs-outdoor. In this way, we force the model to learn a bias, in this case the context. To verify that this model is indeed biased, we perform a set of sanity checks. For that we use an explainability method (GradCAM) on top of mosaics. The nature of mosaics allows us to easily identify the shared bias found within the model: the vegetation patterns were learnt by the model as characteristic of the dog class, while brown and wood patterns are learnt as characteristic of cat. We use this *biased model* to analyze the behavior of the Focus when applied to the biased setting. For baseline we use a *non-biased model*. To perform this experiment, we use 4 mosaic sets: cat-indoor *vs* dog-outdoor (Set 1), cat-indoor *vs* dog-indoor (Set 2), cat-outdoor *vs* dog-indoor (Set 3) and cat-outdoor *vs* dog-outdoor (Set 4). Our findings show how the presence of a shared bias is clearly reflected in the Focus distribution. The Focus decreases when the context learnt by the model is present in both classes within the mosaics. This shows the potential of the Focus, together with the mosaic structure, for the detection of unwanted biases in datasets and models.

6. Future Work

In this work, we empirically proved how the Focus is sensitive to the presence of bias in the model. However, it remains as future work to design a methodology to construct a reduced set of mosaics (selecting a pair of images) that highlight potential biases in the model. The key idea would be to provide to the domain experts the smallest number of mosaics containing as many model biases as possible. Domain experts would use such set of mosaics for the subsequent inspection, detection and classification between desirable and undesirable biases.

Our current method to select mosaics consist in selecting images with lowest prediction of its corresponding class (see §3.3) with the idea of containing characteristic patterns of the other class. However, in a non-binary problem, an image with low prediction score for a class may contain patterns from any of the other resting classes. This casuistic multiplies the number of mosaics that will be provided to the domain expert, thus increasing the complexity of their task.

An interesting case would be a mosaic with high accuracy and low Focus. On one side, the high accuracy would mean that the mosaic contains strong evidences of the *target class* while, on the other side, the low Focus score would mean that such evidences are shared between mosaic images, belonging and non-belonging to the *target class*. This remains as future work.

Acknowledgements

This work is supported by the European Union – H2020 Program under the “INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities”, Grant Agreement n.871042, “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” and by the Departament de Recerca i Universitats of the Generalitat de Catalunya under the Industrial Doctorate Grant DI 2018-100.

References

- [1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.
- [2] Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:170603825. 2017.
- [3] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618-26.
- [4] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one. 2015;10(7).
- [5] Ribeiro MT, 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; 2016. p. 1135-44.
- [6] Zhang J, Bagal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. International Journal of Computer Vision. 2018;126(10):1084-102.
- [7] Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems. 2016;28(11):2660-73.
- [8] Arias-Duart A, Parés F, Garcia-Gasulla D, Gimenez-Abalos V. Focus! Rating XAI Methods and Finding Biases. arXiv; 2021. Available from: <https://arxiv.org/abs/2109.15035>.
- [9] Rao S, Böhle M, Schiele B. Towards Better Understanding Attribution Methods. In: 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. .
- [10] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications. 2019;10(1):1-8.
- [11] Liang W, Zou J. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In: International Conference on Learning Representations; 2021. .
- [12] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer; 2014. p. 740-55.
- [13] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision. 2017;123(1):32-73.
- [14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.
- [15] Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. arXiv preprint arXiv:190409237. 2019.

An Ethical Conversational Agent to Respectfully Conduct In-Game Surveys

Eric ROSELLÓ-MARÍN ^a, Maite LOPEZ-SANCHEZ ^{a,1}, Inmaculada RODRÍGUEZ ^a,
Manel RODRÍGUEZ-SOTO ^b and Juan A. RODRÍGUEZ-AGUILAR ^b

^aDepartment de Matemàtiques i Informàtica, Universitat de Barcelona (UB)

^bArtificial Intelligence Research Institute (IIIA-CSIC)

Abstract. The improvement of videogames highly relies on feedback, usually gathered through UX questionnaires performed after playing. However, users may not remember all the details. This paper proposes an ethical conversational agent, endowed with the moral value of respect, that interacts with the user to perform a survey during the game session. To do so, we use reinforcement learning and the ethical embedding algorithm to ensure that the agent learns to be respectful (i.e., avoid gameplay interruptions) while pursuing its individual objective of asking questions. The novelty is twofold: firstly, the application of ethical embedding outside toy problems; and secondly, the enrichment of a survey oriented conversational agent with this moral value of respect. Results showcase how our ethical conversational bot manages to avoid disturbing user's engagement while getting even a higher percentage of valid answers than a non-ethically enriched chatbot.

Keywords. Machine ethics, Reinforcement Learning, Conversational Agents, User Experience Questionnaires, Video Games

1. Introduction

Human Computer Interaction (HCI) and User eXperience design are fast evolving fields that pursue to improve the design of interactive systems [11]. In the context of UX empirical studies, questionnaires [13] have proven to be useful tools for assessing the user experience of using any computer application, and video games and virtual reality experiences are no exception. Thus, game designers resort to playtesting, which usually is conducted by first letting users play the game, and afterwards, once the playing session has concluded, asking questions about their playing experience [12].

However, users may not remember all details by the end of the experience and, if the number of questions is large, they may lead to user boredom or even user fatigue [25], which hinders the quality of the gathered feedback. Moreover, this disadvantage is aggravated when transitioning back to reality to perform a survey about a Virtual Reality (VR) experience, which can lead to systematic bias as the user is no longer immersed in the virtual world [1].

¹Corresponding Author: maite_lopez@ub.edu. Funded by CI-SUSTAIN (Grant PID2019-104156GB-I00), Crowd4SDG (H2020-872944), COREDEM (H2020-785907), Barcelona City Council through the Fundació Solidaritat UB (code 21S01802-001). Manel Rodriguez-Soto was funded by the Spanish Government with an FPU grant (ref. FPU18/03387).

Conversational agents –interactive systems (embodied or not) that engage in conversation with the user [8]–, offer a new way to collect information, allowing to substitute a traditional survey with an agent that prompts the questions to the user. Indeed, conversational agents have shown to be effective for this task, as they increase both user’s commitment with the survey and the quality of the information elicited [10].

Against this background, we propose to introduce a conversational agent that conducts the survey in-game, as part of the game experience, with the aim of avoiding the detrimental effects of post-game questionnaires, and to ease participation by allowing to stay closer to the context of an ongoing exposure [17]. Nevertheless, this has also the risk of disturbing the game flow [24] if the chatbot does not properly identify when to prompt the user, or even result in the abandonment of the interview due to the player’s cognitive overload [10]. Therefore, we argue that the conversational agent should be respectful with the user’s engagement, and thus, we propose to embed the chatbot with a moral value of *respect*, which should guide the agent to perform the questionnaire without disturbing the user experience.

As social interactions must be considered when designing artificial agents [5], it is becoming apparent that agents’ behaviour should align to human values [2]. Unfortunately, although machine ethics [27,28] is an active research area, very little literature is found on alignment of ethical principles in conversational agents. Some discussions highlighted the need to furnish conversational agents with ethical awareness [7]. However, inducing an ethical behaviour requires some learning, since identifying at design time all situations where this may be required constitutes a complex task.

Our proposal ensures the conversational agent learns to behave ethically by applying ethical embedding, a reinforcement learning approach (see e.g., [18]). This methodology for instilling moral value alignment is founded in the framework of Multi-Objective Reinforcement Learning [20] and the philosophical consideration of values [3] as ethical principles that discern good from bad, and express what ought to be promoted. Examples of human values² include fairness, respect, freedom, security, or prosperity [9].

In particular, our proposal redesigns the conversational agent’s learning environment so that it is ensured that the agent learns to pursue its individual objective of asking as many questions as possible while fulfilling the ethical objective of being respectful with the user’s engagement. This advances the state of the art as it showcases the application of the ethical embedding method beyond toy problems and enriches current survey oriented conversational agents with this moral value of *respect*.

2. Problem Formulation and Scenario

Intuitively, our problem is that of designing an ethical conversational agent that performs in-game surveys. Briefly, we tackle this problem by transforming the learning environment of this agent so that it is guaranteed that the agent learns to be respectful with a user playing the game while eliciting as much player feedback as possible. The learning environment for the conversational agent is a (Multi-Objective) Markov Decision Process (see Subsection 3.1) specified based on the game being played, which in this case is a Pong game played by a simulated user. In this context, we understand respect as not

²Sociology and Psychology have also extensively studied human values, which are often defined as abstract ideals that guide people’s behaviour [23].

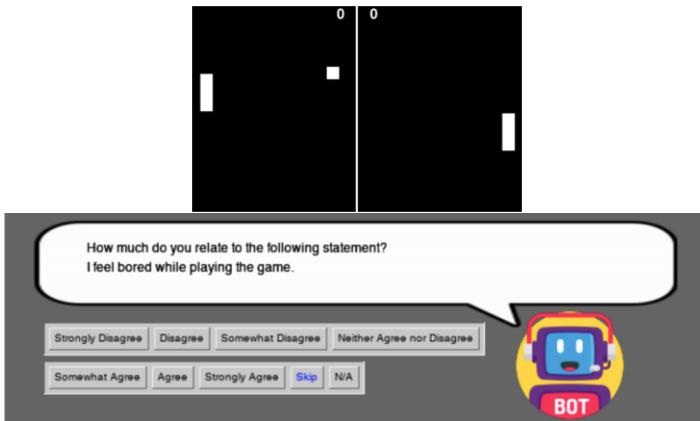


Figure 1. Screenshot of our Pong game illustrating an in-game period in which the chatbot is asking a question (resources from Flaticon, by Freepik). Skip and N/A response options are considered non-valid answers.

hindering the user engagement. In what follows, we introduce engagement and all other necessary elements that characterise our problem scenario.

2.1. Engagement

Within Human-Computer Interaction, engagement is a multi-stage process that becomes key to adapt the designs to the user [16]. The different stages of engagement can be distinguished by different levels of intensity of attributes [15] which, in video games mainly correspond to challenge, aesthetic, feedback, novelty and interactivity.

We can distinguish five different engagement stages. First, the *point of engagement*, is the stage where the user's attention is captured. Next, the *period of engagement* lasts while the attention and interest is maintained through feedback, novelty or challenge. Then, *disengagement* can be followed by the stage of *re-engagement*, which closes the cycle, or *nonengagement*, if the user engagement comes to an end.

In general, as game sessions consist on multiple engagement cycles of varying intensity, we require the survey conversational agent to behave respectful with the user by avoiding interrupting the user engagement, that is, just asking questions when the intensity of the engagement attributes is low.

2.2. Interaction with the User

For the sake of simplicity, we have chosen a single-player three-level Pong game. Levels in this game feature table-tennis games and are interleaved with several transition menus greeting the user or showing the score at the end of each level. Figure 1 depicts an in-game period, where the player uses keyboard arrow keys to move vertically the paddle and hit the bouncing ball. These in-game periods will be the ones typically having high user engagement, as they challenge the users and require from them higher interactivity than menus.

As Figure 1 shows, the conversational agent remains visible at the bottom of the screen throughout the whole game experience, and can prompt questions to the user at any time. Questions are taken from a short version of the Game User Experience

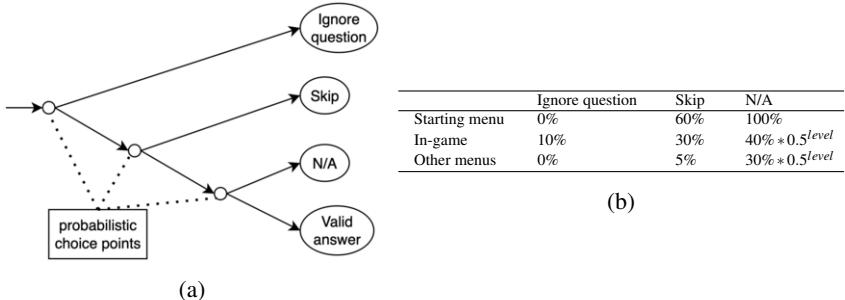


Figure 2. Model of our simulated user, illustrating (a) the rule tree that dictates behaviour and (b) the threshold values of the probabilistic choice points for different in-game or in menu situations.

Satisfaction Scale (GUESS), the GUESS-18, which was designed to be used in iterative game design, testing, and research [12]. Our chatbot asks questions from a pool of 12 questions about enjoyment (see Figure 1), usability/playability, visual aesthetics, etc., discarding those about narrative, audio and social connectivity that do not apply to Pong.

The user can answer any of these questions by selecting the corresponding button in the user interface (see Fig 1). We distinguish two types of answers: *valid* and *non-valid*. Valid answers belong to the Likert scale used in GUESS-18 and are the ones the chatbot should gather to elicit useful data about the user's game experience. Non-valid answers correspond to "Skip" and "N/A": *Skip* denotes the user is not willing to answer a specific question, and thus it is discarded from the pool before being answered; and *N/A* (as in Not Available), indicates the user does not know the answer to the question yet, and should be asked at a later time, so the chatbot still has the chance to get a valid answer later.

Moreover, notice that the player also has the option of ignoring the survey question by simply continue playing. This leaves the chatbot waiting for an answer without being able to pose more questions and without requiring any particular action from the user.

2.3. Simulated user

As previously mentioned, we propose our survey conversational agent to learn to be respectful with the user by applying Reinforcement Learning (RL) [26] methods. However, RL constitutes a data-hungry approach, requiring numerous episodes to learn a policy, and human trials are expensive and time-consuming. Therefore, the repeatability and the acquisition of participants pose a serious challenge [6]. In this context, automatic user simulation tools [21] have been proposed as a handy alternative [14] for the first stages of agents' training, as they provide flexibility and repeatability [21]. Alternative simulators have been proposed based on probabilistic, heuristic, or stochastic models (or a combination of them) [6].

Following heuristic approaches [6] implemented by means of hierarchical patterns (such as HAMs) and rule sets, we have built a simulated user that reproduces human interactions by applying the rule tree in Figure 2a. Non-terminal nodes in the binary tree represent probabilistic *choice points* [22], and terminal nodes indicate the action to be taken. Whenever the chatbot asks a question, the simulated user traverses the tree to decide its reaction. Thus, the probabilities associated to choice point nodes, which

are shown in Figure 2b, allow the random selection of the outgoing edge (i.e., children) to follow. These probabilities vary if the user is playing or not (i.e., in-game or in a menu). We consider the user is collaborative and thus, it never ignores questions while being in a menu (i.e., the “Ignore question” branch in Figure 2 has 0% probability of being selected by the simulated user in Starting menu and Other menus) and just does it 10% in-game (which means it will select any other branch 90% of the times). Overall, we set the probabilities in Figure 2b so that the simulated user will be more likely to provide non-valid answers in-game (i.e., while playing) and in the starting menu than in subsequent menus. Moreover, the further the player gets in the game, the less chances of providing N/A answers. We include these probabilities in order to allow a degree of *lifelike* randomness in the behaviour [14].

3. Background

As previously introduced, we study how a conversational agent can learn to be respectful to the user while performing in-game surveys. The agent’s environment is initially specified as a Multi-Objective Markov Decision Process, which in our approach we transform into a (single-objective) Markov Decision Process. This simplification of the environment is due to the fact that it is simpler for the agent to learn in a single-objective MDP, and thus, it is here where the agent learns its behaviour. Furthermore, we create such single-objective environment in a way that guarantees that the agent will learn a value-aligned behaviour (i.e., policy). This section is devoted to provide the necessary background to introduce our approach.

3.1. Markov Decision Process and Multi-Objective Markov Decision Process

In the context of Reinforcement Learning [26], the learning environment is characterised differently depending on the number of the agent’s learning objectives:

Definition 1. A (single-objective) Markov Decision Process (MDP) is defined as a tuple $\langle S, A, R, T \rangle$ where S is a set of environment states, $A(s)$ is the set of agent actions available at state s , $R(s, a, s')$ is a reward function specifying the reward the agent receives for performing action a at state s when the next state is s' , and $T(s, a, s')$ is the function specifying the probability of such transition.

Definition 2. An n -objective Markov Decision Process (MOMDP) is defined as a tuple $\langle S, A, \vec{R}, T \rangle$ where S , A and T are as in an MDP, and $\vec{R} = (R_1, \dots, R_n)$ is a vectorial reward function composed of n scalar reward functions R_i , one per objective i .

The agent’s behaviour in an (MO)MDP is then described by a policy π , which indicates for each state-action pair $\langle s, a \rangle$, the probability of performing action a in state s . Moreover, a value vector \vec{V} evaluates a policy π by computing the expected discounted sum of rewards obtained when following it:

$$\vec{V}^\pi(s) \doteq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \vec{r}_{t+k+1} | S_t = s, \pi\right] \text{ for every state } s \in S, \quad (1)$$

where $\gamma \in [0, 1]$ is the discount factor and t is the time-step of each state s . An *optimal policy* in a single-objective MDP is, then, one that maximises the expected discounted reward accumulation for every state ($\pi_* \doteq \arg \max_\pi V^\pi$). π_* constitutes the behaviour the

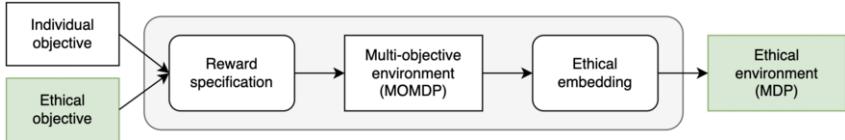


Figure 3. The ethical environment design process (as in [19]) for value alignment.

agent should learn, or, in other words, the solution to the MDP. Its computation is more complex for an MOMDP though, as it involves the optimisation of the value vector \vec{V}^* instead of a single V^* value function.

3.2. Value Alignment

MOMDPs facilitate learning value-aligned behaviours, as they can be used to design the environment to incentivize ethical behaviour. Following the approach in [19], Figure 3 illustrates value alignment as a process consisting of two steps: *reward specification* and *ethical embedding*.

Firstly, the reward specification defines an MOMDP by considering both the individual objective (the agent's original objective translated into individual reward R_0) and the ethical objective (the moral value we introduce). This ethical objective encodes the moral value into rewards and is composed of two dimensions: the *normative* reward function R_N , which punishes the violation of normative moral requirements; and the *evaluative* reward function R_E , which rewards morally praiseworthy actions. In this context, we follow [19] and consider an *ethical policy* as one that abides to all norms while behaving as praiseworthy as possible, and an *ethical-optimal policy* as one that maximizes the individual objective as much as possible subject to being ethical. Formally, we refer to this value-enriched MOMDP as an *ethical* MOMDP, and define it as $\langle S, A, (R_0, R_N + R_E), T \rangle$.

Secondly, Figure 3 (right) depicts how the ethical embedding process transforms this ethical MOMDP into a single-objective MDP, where the agent is incentivized to learn an *ethical-optimal policy*. That is, the resulting MDP guarantees that the agent learns to fulfil the ethical objective while pursuing its individual objective (and, as it is single-objective, just requires the agent to apply a basic reinforcement learning method).

The ethical embedding process applies this transformation by computing a linear scalarisation function over the vectorial rewards \vec{R} in the MOMDP that results in a scalar reward function R for an ethical MDP. This function has the form of:

$$f(\vec{V}^\pi) = \vec{w} \cdot \vec{V}^\pi = w_0 V_0^\pi + w_e (V_N^\pi + V_E^\pi) \quad (2)$$

Following [19], we fix the individual weight $w_0 = 1$ so that the ethical embedding process is reduced to looking for the ethical weight $w_e > 0$ that guarantees the learned behaviour in the resulting ethical MDP $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$ will prioritise the ethical objective over the individual one.

Algorithm 1 illustrates this computation. First, it applies Convex Hull Value Iteration [4], a modification of the original Bellman's Value Iteration algorithm [26] that allows learning the optimal policies for all linear preference assignments over multiple objectives. The resulting convex hull contains the subset of policies that are optimal for some value of the ethical weight w_e . Thus, second line of the algorithm exploits the convex hull to extract from it the value of the policy with the maximum amount of ethical value

$(V_N + V_E)$ (i.e., the value \vec{V}^* of the ethical-optimal policy π^*), and the value of the policy with the second-best value (\vec{V}'^*) . Next, third line finds the values of w_e for which the former policy becomes optimal by computing the minimal weight satisfying:

$$V_0^*(s) + w_e[V_N^*(s) + V_E^*(s)] > V'_0(s) + w_e[V'_N(s) + V'_E(s)]. \quad (3)$$

Algorithm 1 Ethical Embedding [19]

```

function EMBEDDING( Ethical MOMDP  $\langle S, A, (R_0, R_N + R_E), T \rangle$ )
    Compute the convex hull for weight vectors  $\vec{w} = (1, w_e)$  with  $w_e > 0$ 
    Find  $\vec{V}^*$  the ethical-optimal value vector, and  $\vec{V}'^*$  the second-best value vector in the convex hull
    Find the minimal value for  $w_e$  that satisfies Eq. 3
return  $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$ 
```

4. Environment design for an in-game survey agent to learn to be respectful

As previously mentioned, the ethical environment design process first defines an ethical MOMDP to then transform it into an ethical MDP by applying the embedding algorithm.

In our particular setting (see Figure 2), we define our ethical MOMDP $\langle S, A, \vec{R}, T \rangle$ so that states in S include information about current game status (level and if menu or in-game) and user's activity (if engaged³ or if the answer to last question was valid/non-valid or quick/slow). Moreover, the agent can perform two actions $A = \{\text{Ask}, \text{Wait}\}$ and the reward vector $\vec{R} = (R_0, R_N + R_E)$ contains the individual and ethical reward functions:

- R_0 (individual reward): promotes collecting as many valid answers as possible.

$$R_0(s, a, s') \doteq \begin{cases} 1, & \text{if } a = \text{Ask and } \text{valid_answer}(s') \\ 0, & \text{otherwise} \end{cases}$$

- R_N (normative reward): punishes i) asking questions when the user is engaged or provides non-valid or slow answers; and ii) waiting (i.e., not asking questions) when the user is not engaged, as these moments of low engagement should not be wasted:

$$R_N(s, a, s') \doteq \begin{cases} -2, & \text{if } ((a = \text{Ask and } (\text{engaged}(s) \text{ or not } \text{valid_answer}(s') \text{ or } \text{slow_answer}(s')) \\ & \quad \text{or } (a = \text{Wait and not } \text{engaged}(s))) \\ 0, & \text{otherwise} \end{cases}$$

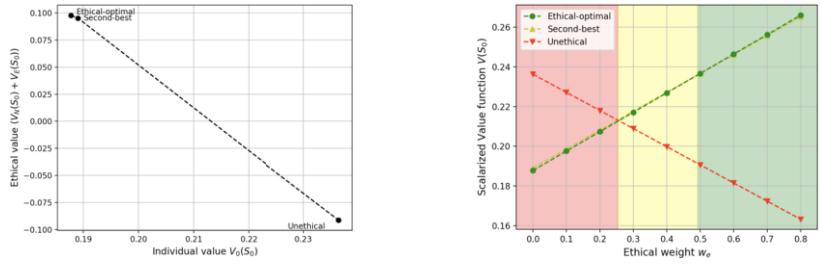
- R_E (evaluative reward): promotes asking questions that get a quick and valid response without interrupting engagement:

$$R_E(s, a, s') \doteq \begin{cases} 1, & \text{if } (a = \text{Ask and } \text{quick_answer}(s') \text{ and } \text{valid_answer}(s') \text{ and not } \text{engaged}(s)) \\ 0, & \text{otherwise} \end{cases}$$

Thus $R_N + R_E$ encapsulates our notion of respect applied to the context of performing in-game questionnaires. Finally, state transition probabilities in $T(s, a, s')$ are approximated by observing the frequencies of such transitions in 500 game executions.

Next, we apply the ethical embedding algorithm. Figure 4a visualizes the convex hull, that is, those policies that are maximal for some value of w_e . Specifically, black dots signal the ethical-optimal policy (\vec{V}^*) , the one that maximizes the ethical value function

³Notice that in our simple Pong game, engagement can be assumed if the user moves the paddle, but this varies for different games. Moreover, although moving the paddle can only be done in-game, and thus we assume low engagement in menus, it may also happen if the play is slow enough.



(a) The convex hull for our ethical MOMDP.

(b) Scalarized policies in the weight space.

Figure 4. The ethical embedding process: (a) visualizing the convex hull, and (b) finding the ethical weight.

$(V_N + V_E)$); the second-best ethical optimal policy (\vec{V}'^*); as well as the (unethical) policy that maximizes the individual value (V_0). Next, we solve Eq. 3 and obtain a value of $w_e > 0.49237$. In fact, this value can be empirically found by plotting, as in Figure 4b, the scalarised values for these three policies, and by identifying the value of w_e for which the ethical-optimal policy has the highest scalarised value (and this is also the case for all w_e values in the green area). Then, we set the weight to $w_e = 0.5$ and return the ethical MDP $\langle S, A, R_0 + w_e(R_N + R_E), T \rangle$ as the environment that guarantees that the agent will learn to behave ethically. Finally, it is worth mentioning that Theorem 1 in [19] formally guarantees that the agent will still learn the same ethical optimal policy regardless of the scale⁴ of the ethical rewards considered before scalarisation.

5. Results

The resulting ethical MDP provides a simple environment for our conversational agent to learn to be respectful while asking survey questions. Here, we empirically prove so by applying Q-learning [26]. Specifically, we set a learning rate $\alpha = 0.7$, a discount factor $\gamma = 0.7$, and an ϵ -greedy policy for exploration along 1000 episodes, where each episode corresponds to a playthrough of our three-level Pong game⁵.

To better assess the impact of the ethical embedding. Figure 5a illustrates the convergence, in terms of the accumulated reward, of the learning of two agents: in green, our ethical agent; in red, an unethical agent that just considers the individual reward R_0 . Not surprisingly, our ethical agent takes longer to learn, and accumulates negative rewards as the R_N reward is quite demanding and punishes the agent for not taking advantage of all low engagement situations in slow play. However, this does not preclude our ethical agent to elicit necessary information. In fact, as depicted in Figure 5b, once it learns, it manages to get more valid answers than the unethical agent, which relies on the user to answer questions even if interrupted.

Beyond checking that the ethical agent manages to accomplish its individual objective, we need to assess it learns a respectful behaviour, asking questions when the user's engagement is low, which typically happens while the user is in menus. Thus, we focus on comparing the number of questions prompted in-game and in menus. Specifically, Figure 5c shows how the green ethical agent manages to drastically reduce the number of questions in-game (as opposed to the red unethical agent) and Figure 5d shows how the

⁴As long as the reward of praiseworthy actions are > 0 and the ones for blameworthy actions are < 0 .

⁵Our code is publicly available at <https://github.com/ericRosello/EthicalCA>.

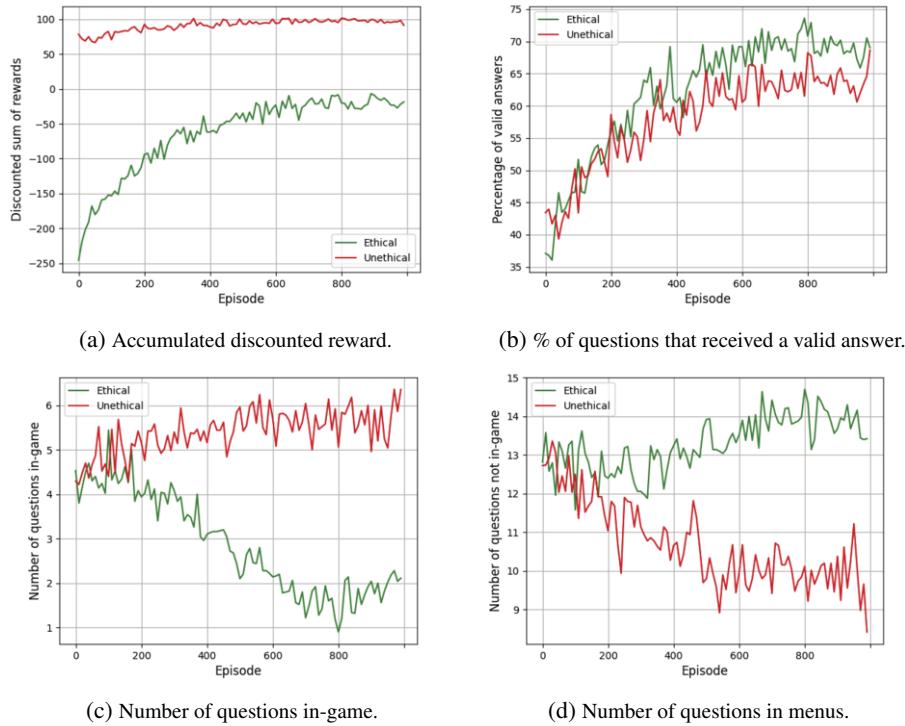


Figure 5. Evolution of different metrics throughout the learning process.

ethical agent focuses in asking most of the questions in menus (a behaviour that again contrasts with the one of the unethical agent). Thus, overall, we can claim that our conversational agent has successfully learnt to ask survey questions without disturbing the user play, that is, behaving in alignment with the moral value of respect.

6. Conclusions and Future Work

This paper proposes an ethical conversational agent in charge of gathering User eXperience data while the user is playing a game. The agent, applying the ethical embedding method, learns to respectfully conduct the in-game questionnaire. This method transforms an ethical MOMDP into an ethical MDP that can be addressed by standard RL algorithms. Specifically, we defined the learning environment based on the Pong game, and used Q-learning with a simulated user to assess the ethical agent's learning. The results show that our ethical agent asks the user questions in more appropriate situations (low user engagement) than the unethical agent. Thus, it fulfils the ethical objective while still pursuing the individual one (i.e. obtain as much UX data as possible). Indeed, the ethical agent obtained a higher proportion of valid answers than the unethical one, while reducing gameplay interruptions.

Future work should explore the generalization of our approach to alternative games and virtual reality experiences, as the activity of the user (and so engagement) is highly dependent on the (game) mechanics. The study of other moral values (e.g. fairness) is another interesting line of research.

References

- [1] D. Alexandrovsky, S. Putze, M. Bonfert, S. Höffner, P. Michelmann, D. Wenig, R. Malaka, and J. D. Smeddinck. Examining design choices of questionnaires in vr user studies. In *CHI'20*, 1–21, 2020.
- [2] M. Anderson, S. L. Anderson, and C. Armen. An approach to computing ethics. *IEEE Intelligent Systems*, 21(4):56–63, 2006.
- [3] T. Arnold, D. Kasenberg, and M. Scheutz. Value alignment or misalignment - what will keep systems accountable? In *AAAI Workshops*, 2017.
- [4] L. Barrett and S. Narayanan. Learning all optimal policies with multiple criteria. *Proc. of 25th ICML*, pages 41–47, 01 2008.
- [5] A. Beck, B. Stevens, K. A Bard, and L. Cañamero. Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):1–29, 2012.
- [6] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale. An evaluation methodology for interactive reinforcement learning with simulated users. *Biomimetics*, 6(1):13, 2021.
- [7] J. Casas-Roma and J. Conesa. Towards the design of ethically-aware pedagogical conversational agents. In *Int. Conference on 3PGCIC*, pages 188–198. Springer, 2020.
- [8] J. Cassell. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, 22(4):67–67, 2001.
- [9] A. Cheng and K. R. Fleischmann. Developing a meta-inventory of human values. *Proc. of the ASIS&T*, 47(1):1–10, 2010.
- [10] X. Han, M. Zhou, M. J. Turner, and T. Yeh. Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proc. of CHI'21*, pages 1–15, 2021.
- [11] M. Hassenzahl. User experience and experience design. *The encyclopedia of HCI*, 2, 2013.
- [12] J. R. Keebler, W. J. Shelstad, D. C. Smith, B. S. Chaparro, and M. H. Phan. Validation of the guess-18: a short version of the game user experience satisfaction scale. *J. of Usability Studies*, 16(1):49, 2020.
- [13] E. L. Law. The measurability and predictability of user experience. In *ACM SIGCHI EICS*, 1–10, 2011.
- [14] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23, 2000.
- [15] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the ASIS&T*, 59(6):938–955, 2008.
- [16] C. Peters, G. Castellano, and S. De Freitas. An exploration of user engagement in hci. In *Proc. of the Int. Workshop on Affective-Aware Virtual Agents and Social Robots*, pages 1–3, 2009.
- [17] I. Rodríguez and A. Puig. Open the microphone, please! conversational ux evaluation in virtual reality. In *Workshop 'Evaluating user experiences in mixed reality' in CHI'21*, 2021.
- [18] M. Rodriguez-Soto, M. Serramia, M. Lopez-Sánchez, and J. A. Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):1–17, 2022.
- [19] Manel Rodriguez-Soto, Maite Lopez-Sánchez, and Juan A Rodriguez-Aguilar. Multi-objective reinforcement learning for designing ethical environments. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 1–7, 2021.
- [20] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, October 2013.
- [21] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126, 2006.
- [22] K. Scheffler and S. Young. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proc. of HLT*, volume 2, 2002.
- [23] S. H. Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.
- [24] V. J. Shute. Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2):503–524, 2011.
- [25] A. Steinmauer, M. Sackl, and C. Gütl. Engagement in in-game questionnaires-perspectives from users and experts. In *2021 7th Int. Conference of the iLRN*, pages 1–7. IEEE, 2021.
- [26] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [27] W. Wallach and C. Allen. *Moral machines: teaching robots right from wrong*. Oxford Univ. press, 2008.
- [28] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang. Building ethics into artificial intelligence. In *IJCAI*, page 5527–5533, 2018.

Fuzzy-LORE: A Method for Extracting Local and Counterfactual Explanations Using Fuzzy Decision Trees

Najlaa MAAROOF^{a,1}, Antonio MORENO^a Mohammed JABREEL^a and Aida VALLS^a

^a ITAKA-Intelligent Technologies for Advanced Knowledge Acquisition - Departament d'Enginyeria Informàtica i Matemàtiques Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

Abstract. Classification systems based Machine Learning hide the logic of their internal decision processes from the users. Hence, post-hoc explanations about their predictions are often required. This paper proposes Fuzzy-LORE, a method that generates local explanations for fuzzy-based Machine Learning systems. First, it learns a local fuzzy decision tree using a set of synthetic neighbours from the input instance. Then, it extracts from the logic of the fuzzy decision tree a meaningful explanation consisting of a set of decision rules (which explain the reasons behind the decision), a set of counterfactual rules (which inform of small changes in the instance's features that would lead to a different outcome), and finally a set of specific counterfactual examples. Our experiments on a real-world medical dataset show that Fuzzy-LORE outperforms prior approaches and methods for generating local explanations.

Keywords. Explainable AI (XAI), Machine Learning, Fuzzy Decision Tree, Diabetic Retinopathy, LORE

1. Introduction

Machine Learning (ML)-based systems have become a vital component of multiple applications in many domains, especially in healthcare. One key reason for their widespread adoption is the success in the development of accurate classification models, which help doctors in the diagnosis, treatment and prognosis of complex diseases. One example of those successful classification methods is the Fuzzy Random Forest (FRF), which can be used to solve multi-class or binary classification problems. A FRF is composed of hundreds of Fuzzy Decision Trees (FDTs) [1].

Despite their accuracy, most modern ML-based systems are considered black boxes, because it is not straightforward to understand the reasons behind their

¹Corresponding Author: Najlaa Maaroof. E-mail: najlaamaaroofwahib.al-ziyadi@urv.cat

decisions. As a result, developing methods for explaining them has become highly demanded [2, 3].

Several kinds of explanation methods have been proposed in the literature. A popular approach is to use post-hoc explanation methods, which study the relationship between the input and the output produced by the system to extract a local explanation of a particular decision on an instance x . Most of the methods that follow this approach generate a set of inputs (neighbours of x), analyse the answers provided by the system to be explained and then create a simpler model from which a local explanation can be inferred [4, 5].

One of the most well-known post-hoc explanation methods is *Local Rule-Based Explanations* (LORE, [6]), explained briefly in section 2. LORE derives a rule-based explanation composed of the activated rule used to explain the rationale behind the system's decision, and a set of counterfactual rules which represent the minimal number of changes in the feature values of the instance that would change the conclusion of the system. Such counterfactual explanation is useful in domains like healthcare. It helps practitioners to decide what they should do to obtain a desired state instead of providing them only with important features that led to the decision.

Although LORE has shown a good performance in explaining classical ML-based systems [6], we believe that it can be improved for the particular case of fuzzy-based systems. In our previous works [7, 8] we proposed two extensions of LORE, called *Guided-LORE* and *C-LORE-F*. In the former the neighbours' generation step was formalized as a search problem and solved using Uniform Cost Search, whereas in the latter the knowledge about the definition of the fuzzy variables was used to focus the exploration of the neighbours' space. Such adaptations allowed us to make the generation process more informed and leverage more contextual information, mainly in the case in which the attributes that define the objects are fuzzy, covered in *C-LORE-F*.

Despite the promising outcome obtained with Guided-LORE and C-LORE-F, they still have some shortcomings. First, the quality of the obtained counterfactual instances should be improved [7]. Second, the basic explanation in LORE (and its variants) is limited to a single rule derived from the activated path in a decision tree, which is not very informative. Third, the method is quite rigid and the explanation can't be adapted to different applications or user types.

In this work, we propose a novel method called Fuzzy-LORE to address the shortcomings of standard LORE-based methods (i.e., LORE, Guided-LORE and C-LORE-F) and provide better explanations in the case of fuzzy-based ML systems. Fuzzy-LORE adapts our previous LORE-based methods by using fuzzy decision trees as an alternative to the classical decision trees. First, it learns a local fuzzy decision tree predictor on a synthetic neighbourhood of the instance x to be explained. Then, it extracts from the logic of the fuzzy decision tree a meaningful explanation consisting of a set of decision rules, a set of counterfactual rules, and a set of counterfactual examples. We will focus only on binary classification.

We evaluated the proposed method on a private dataset, used to train a FRF-based binary classifier that assesses the risk of developing diabetic retinopathy in diabetic patients. The experimental results show that, according to several met-

rics, Fuzzy-LORE outperforms the prior classical LORE-based methods, mainly in the generation of counterfactual examples.

The rest of this article is structured as follows. Section 2 provides an overview of the classical LORE-based methods. Section 3 explains the proposed method. In Section 4, we describe the experimental setup and discuss the obtained results. Finally, in section 5, we conclude the paper and list some points for future work.

2. Preliminaries

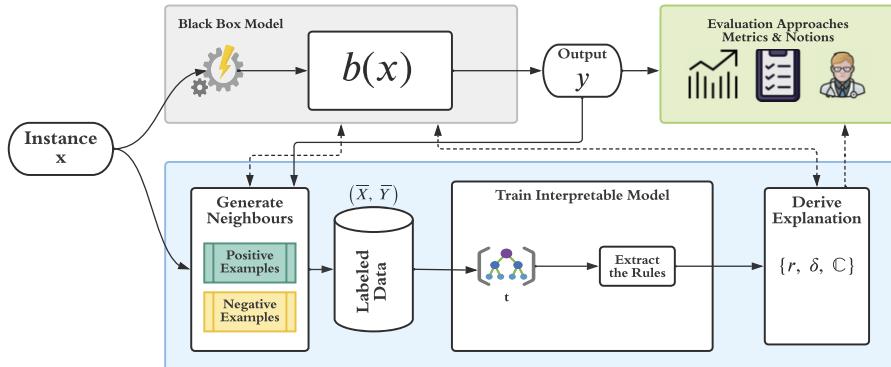


Figure 1. Architecture of the LORE-based explanation methods.

This section provides a brief background of the classical LORE-based methods, which use decision trees to provide a post-hoc explanation for the decision assigned to a specific instance. The inputs of the LORE-based method are a trained ML model, b , and an example x . Figure 1 shows the main steps and the general architecture of LORE. First, b is applied to x to get a decision y . Then, we obtain a set of neighbours of x , \mathcal{D} , and a rule-based model t (a decision tree) is built by considering the output of b in these points. From this model t it is possible to derive an explanation that contains the rule r used to classify x , a set of counterfactual rules δ and a set of counterfactual instances \mathbb{C} .

The set \mathcal{D} is obtained by merging two subsets, \mathcal{D}^+ and \mathcal{D}^- . The first one is called the *positive set*, and it contains a set of instances that belong to the same class than x . The second one, the *negative set*, contains examples with a different class. We obtain \mathcal{D}^- by looking at an auxiliary set T and finding the closest example to x , i.e., x^- , that has a different label than y . T can be the training set used to train the black-box model b , if accessible, or any other data set from the same distribution.

The main difference between LORE and its extensions Guided-LORE and C-LORE-F, lies in the neighbours' generation step. LORE uses a genetic algorithm to do it, whereas our methods define the neighbourhood generation as a search problem in which we explore the neighbourhood space of a point x by applying a Uniform Cost Search.

Once the synthetic neighbours, i.e., \mathcal{D} have been obtained, a decision tree predictor, t , is trained on \mathcal{D} . Finally, the explanation is derived from t by finding the activated path and converting it to a rule format. Then, the counterfactual rules and their corresponding counterfactual instances are extracted from the analysis of t .

As stated in the introduction, the replacement of the decision tree by a fuzzy decision tree requires changes in the tasks of extracting the decision rules, the counterfactual rules and the counterfactual instances, which are detailed in the next section.

3. Fuzzy-LORE

Fuzzy-LORE is an adaptation of the LORE explanation methods that employs a FDT local predictor to obtain richer linguistic explanations of binary classification systems based on fuzzy variables. Following the LORE scheme, Fuzzy-LORE first generates synthetic neighbours of the instance of interest, using the method proposed in C-LORE-F [8]. This mechanism utilizes contextual information from the definitions of the fuzzy sets of the attributes. Fuzzy-LORE uses fuzzy tools in all of its steps, from generating the neighbours to constructing the FDT local predictor model and obtaining the explanation components. The next subsections explain in detail all the steps of Fuzzy-LORE.

3.1. FDT Construction

The first novelty in Fuzzy-LORE is the construction of a local interpretable model consisting of a small fuzzy decision tree. The fact that the tree uses the same linguistic variables than the fuzzy black-box model will help to interpret the explanations. Fuzzy-LORE uses the induction algorithm proposed in [9] to construct a local fuzzy decision tree with conjunctive rules. This algorithm is an adaptation of ID3 for fuzzy datasets with linguistic variables. It uses two parameters during the construction process. The first one is called the significance level, α , which filters out the evidence that is not relevant enough. The second parameter is the truth level threshold, β , which controls the tree's growth as it defines the minimum level for ending a branch. In the experimental section, the values $\alpha = 0.1$ and $\beta = 0.9$ have been empirically obtained.

The main steps to construct the FDT are the following: (1) Select the best attribute as the root of the tree, based on the ambiguity function [1]. (2) For each linguistic term of this attribute, create a branch with the examples with support of at least α , and compute the truth level of classification for each class in the set of classes. (3) If the truth level of classification is above β for at least one class, terminate the branch and set the label as the class with the highest truth level. (4) Otherwise, check if an additional attribute will further reduce the classification ambiguity. If that is the case, select the best one as a new decision node of the branch and repeat step 2 until no further growth is possible. (5) Otherwise, terminate the branch as a leaf with a label corresponding to the class with the highest truth level. After constructing the tree, each branch can be

considered as a classification rule with a degree of support equal to the truth level of its conclusion.

3.2. Inference in FDT

The Mamdani inference procedure is applied to find the decision class for an input, x , as follows: (1) Calculate the satisfaction level of the premises of each rule, using the t-norm minimum. (2) Calculate the membership of x to the conclusion class as the product between the satisfaction level of the premises and the degree of support of the rule. (3) Aggregate all the memberships for the same class, given by different rules, using the t-conorm maximum. The result is the confidence on the class. (4) For binary problems, compare the confidences of class 0 and class 1 and choose the one that has the highest value as the final decision class.

3.3. Explanation Extraction from FDT

The second change in Fuzzy-LORE is the explanation extraction process. Fuzzy-LORE derives an explanation from the constructed FDT, slightly different from the one derived by the LORE-based methods. Having in mind that we are dealing with binary classification problems, and given that $b(x) = y$, the explanation of this classification has the form of a triplet $(\mathbb{R}, \Delta, \mathbb{C})$, where:

- \mathbb{R} is the set of decision rules that cover the instance x and have y as output. Each rule $r \in \mathbb{R}$ tells which conditions are satisfied by the object x for being classified as y . Thus, they indicate several minimal sets of conjunctive conditions necessary for belonging to that class.
- Δ is the set of counterfactual rules that lead to the opposite class.
- \mathbb{C} is a set of counterfactual instances, that represent examples of objects that do not belong to class y and have the minimum changes with respect to the original input object x .

3.3.1. Decision rules

Let $R_x = R_x^+ \cup R_x^-$ be the set of all fuzzy rules of the constructed FDT given the instance x . R_x^+ refers to the set of rules that have the conclusion y , and R_x^- is the set of fuzzy rules that have the opposite conclusion. Each rule $r \in R_x^+$ has the following format:

$$IF (f_i IS t_{i,a}) AND (f_j IS t_{j,b}) \dots AND (f_z IS t_{z,c}) THEN class IS y$$

Each attribute f_i is a linguistic variable with a set of terms $t_{i,1}, t_{i,2}, \dots$. Each term has an associated fuzzy set, $\mu_{t_{i,a}}$, and they define a fuzzy partition.

R_x^+ contains the FDT rules activated by x that lead to the conclusion y , so they constitute the base of the decision rules of the explanation. In order to present a comprehensible explanation, Fuzzy-LORE is more flexible than LORE, that has one single activated crisp rule. In this new version, \mathbb{R} can be defined as a subset of the rules in R_x^+ , taking advantage of the fuzzy activation of several

rules. Depending on the application and on the user type, \mathbb{R} can be either all the rules in R_x^+ , the top k rules with highest confidence scores, or the set of rules with a confidence above a certain threshold. In the experiments presented in this paper, the top 3 rules with highest confidence were included in \mathbb{R} .

3.3.2. Counterfactual rules

The aim of this step is to find rules similar to those of \mathbb{R} which lead to the opposite conclusion. After finding these counterfactual rules, in the next step it will be possible to compute counterfactual instances (individuals close to x that belong to a different class).

To extract the counterfactual rules Δ , we consider each rule $r_c \in R_x^-$. For each condition $c_i = (f_i \text{ IS } t_{i,a})$ in r_c , we check if f_i appears in any condition of the rules in \mathbb{R} with a different term, i.e. $t_{i,b}$. We change the membership of $t_{i,b}$ by its negation $1 - \mu_{t_{i,b}}(x)$. Then, we re-calculate the final confidence score of the rule r_c with the new membership values as mentioned in subsection 3.2. With the negated membership functions we are analysing what would happen if we changed the original value of x in f_i to a value that activated that condition.

After that, we filter out those rules in R_x^- that have confidence smaller than the maximum confidence score in \mathbb{R} and these are the final counterfactual rules of the explanation component Δ .

3.3.3. Counterfactual instances

Finally, for each rule r_c in Δ we create a counterfactual instance $x_c \in \mathbb{C}$ by making a copy of x and changing only the values of the features that appear in r_c . Concretely, each term t_i of r_c is defuzzified using the Center-of-Maximum method (using the membership values calculated in the previous step). The rationale for substituting the original value by the center of maximum is to put a value in the attribute that maximally activates the condition of the counterfactual rule.

Let us illustrate the process of generating a counterfactual instance with an example from the problem of diagnosis of diabetic retinopathy, used in the experimental section. The instance that is classified is $x = \{\text{Age}=61, \text{Sex}=1, \text{EVOL}=19, \text{TTM}=2, \text{HbA1c}=9, \text{CDKEPI}=106.21, \text{MA}=0, \text{BMI}=44.21, \text{HTAR}=1\}$. For simplification we only consider one decision rule r in R_x^+ .

$$r : \text{IF } (\text{HbA1c IS More9}) \text{ THEN class IS Class1} \text{ (confidence = 0.757)}$$

One of the candidate counterfactual rules is

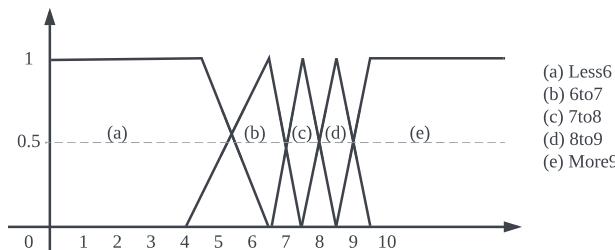


Figure 2. The membership functions of the HbA1c variable.

$r_c : IF (HbA1c \text{ IS } 6\text{to}7) \text{ AND } (MA \text{ IS } \text{Correct}) \text{ THEN class IS Class0}$
with a confidence of 0 based on x .

So, to construct a counterfactual instance only the value of attribute HbA1c must be changed as it appears in both r and r_c . Given the fuzzy set definitions for HbA1c shown in Figure 2, the fuzzification of the value HbA1c=9 of x gives the following membership scores for each term: $\{\text{Less6} = 0, 6\text{to}7 = 0, 7\text{to}8 = 0, 8\text{to}9 = 0.2, \text{More9} = 0.8\}$.

To activate the condition ($HbA1c \text{ IS } 6\text{to}7$) in r_c , we take the negated membership of this term, $\mu_{6\text{to}7} = 1$. Based on that, the confidence of r_c will be 0.78 instead of 0 and it will be included in Δ as its confidence is larger than 0.757.

Finally, the following counterfactual example is obtained: $\{\text{Age}=61, \text{Sex}=1, \text{EVOL}=19, \text{TTM}=2, \text{HbA1c}=6.5, \text{CDKEPI}=106.21, \text{MA}=0, \text{BMI}=44.21, \text{HTAR}=1\}$

4. Experiments and Results

This section describes the experimental setup, and discusses the obtained results by comparing the performance of different methods and evaluating the generated counterfactual examples.

4.1. Experimental Setup

We evaluated Fuzzy-LORE on a private data set that shows if a diabetic patient has (or not) a high risk of developing diabetic retinopathy. It is composed of 2323 examples of binary classification. The Diabetic-Retinopathy data set was used to develop a fuzzy random forest-based classifier, called RETIPROGRAM, which is currently being used in the Hospital de Sant Joan in Reus (Tarragona). Each instance in the data set is defined by nine attributes: current age, sex, years since diabetes detection, type of diabetes treatment, good or bad control of arterial hypertension, HbA1c level, glomerular filtrate rate estimated by the CKD-EPI value, microalbuminuria, and body mass index. The data was split into a training set of 1212 examples and a test set of 1111 examples. The classification model used in RETIPROGRAM achieves an accuracy of 80%, with a sensitivity of 81.3% and a specificity of 79.7% [10]. We used the test set in all our experiments to evaluate the effectiveness of Fuzzy-LORE.

4.2. Evaluation of the Explanation Results

As described in the previous section, a Fuzzy-LORE explanation contains the explanation decision rules \mathbb{R} and a set of counterfactual rules Δ , from which the counterfactual examples, \mathbb{C} , are derived. These components are obtained from a fuzzy decision tree (a set of fuzzy decision rules), that we call the explanation model. In this section we evaluate the quality of the rules generated by the proposed method and compare it to the LORE-based methods using the following evaluation metrics:

- **Hit:** this metric computes the similarity between the output of the explanation model and the black-box, b , for all the testing instances. It returns 1 if they are equal and 0 otherwise.
- **Fidelity:** this metric measures to which extent the explanation model can accurately reproduce the black-box predictor for the particular case of instance x . It answers the question of how good is the explanation model at mimicking the behaviour of the black-box by comparing its predictions and the ones of the black-box on the instances that are neighbours of x , which are in \mathcal{D} .
- **l-Fidelity:** it is similar to the *fidelity*; however, it is computed on the subset of instances from \mathcal{D} covered by the explanation rules, \mathbb{R} . It is used to measure to what extent these rules are good at mimicking the black-box model on similar data of the same class.
- **c-Hit:** this metric compares the predictions of the explanation model and the black-box model on all the counterfactual instances of x , \mathbb{C} .

Table 1 shows the means and standard deviations of the metrics for Fuzzy-LORE and the previous LORE-based methods on the test set. It may be seen that Fuzzy-LORE and C-LORE-F show almost the same performance in the Hit and Fidelity measures. C-LORE-F is slightly better than Fuzzy-LORE in terms of l-Fidelity. However, Fuzzy-LORE outperforms clearly all the other methods in terms of c-Hit. We can attribute such improvement in the c-Hit measure to the quality of the generated counterfactual examples (which are evaluated in more depth in Section 4.3).

Table 1. Evaluation of the explanation results for Fuzzy-LORE vs other LORE-based methods.

Methods	Hit	Fidelity	l-Fidelity	c-Hit
LORE	0.95 ± 0.13	0.96 ± 0.05	0.95 ± 0.09	0.79 ± 0.32
Guided-LORE	0.99 ± 0.02	0.98 ± 0.06	0.99 ± 0.03	0.83 ± 0.28
C-LORE-F	1.00 ± 0.00	0.99 ± 0.002	0.99 ± 0.002	0.89 ± 0.29
Fuzzy-LORE	1.00 ± 0.00	0.99 ± 0.03	0.98 ± 0.04	0.96 ± 0.17

4.3. Evaluation of the counterfactual examples

Counterfactual examples help to understand what changes may be applied to an object to obtain a different outcome. This is particularly interesting in health-care applications. Hence, it is important to have counterfactual examples that balance a wide range of suggested modifications (diversity) and the relative facility of adopting those modifications (proximity to the actual input). Moreover, counterfactual examples must be actionable, e.g., people can not reduce their age or change their race.

In this subsection, we evaluate the generated counterfactual examples for C-LORE-F and Fuzzy-LORE (as they showed almost the same performance) using the following evaluation metrics [11]:

- **Validity:** is the number of counterfactual examples with a different outcome than the original input, i.e., x , divided by the total number of counterfactual examples.

$$\text{Validity} = \frac{|\hat{x} \in \mathbb{C} \text{ s.t. } b(x) \neq b(\hat{x})|}{|\mathbb{C}|} \quad (1)$$

Here \mathbb{C} refers to the set of returned counterfactual examples and b is the black-box model.

- **Proximity:** is the mean of feature-wise normalised distances between a counterfactual example c and the original input x .

$$\text{Proximity} = 1 - \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} \text{dist}(c, x) \quad (2)$$

- **Sparsity:** it measures the average of attribute value changes between a counterfactual example and the original input.

$$\text{Sparsity} = 1 - \frac{1}{|\mathbb{C}| * |\mathbb{F}|} \sum_{c \in \mathbb{C}} \sum_{f \in \mathbb{F}} \mathbb{1}[c_f \neq x_f] \quad (3)$$

Here, \mathbb{F} is the set of features, and $\mathbb{1}$ is the indicator function.

- **Diversity:** it is similar to proximity. However, instead of computing the feature-wise distance between the counterfactual example and the original input, we compute it between each pair of counterfactual examples.

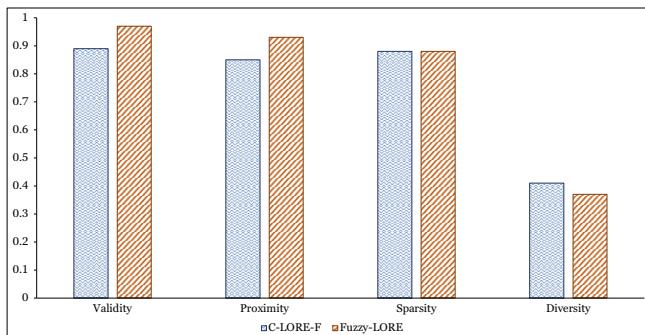


Figure 3. Evaluation of the counterfactual examples for C-LORE-F and Fuzzy-LORE.

Figure 3 shows the comparative results of Fuzzy-LORE vs C-LORE-F with respect to these evaluation metrics. In general, Fuzzy-LORE showed better performance than C-LORE-F, mainly in terms of validity and proximity. Both Fuzzy-LORE and C-LORE-F have similar performance in terms of sparsity. Looking at the diversity results, we can find that C-LORE-F generates slightly more diverse counterfactual examples than the proposed method. However, both of them showed a low performance. This issue will be studied in future work.

5. Conclusion

Fuzzy-LORE is a new post-hoc explanation method for fuzzy binary classifiers. It learns a local fuzzy decision tree on a synthetic neighbourhood of an instance. Then, it extracts from it a meaningful explanation consisting of : (1) A set of decision rules that explain the reasons behind the classification decision. (2) A set of counterfactual rules that suggest a minimal number of changes in the instance features to get a different outcome. (3) A set of counterfactual examples. The method has been evaluated on a dataset to assess the risk of developing diabetic retinopathy. The evaluation results revealed that using the fuzzy decision tree as an explanation model gives better explanations than the decision tree, mainly in the counterfactual rules and instances. However, Fuzzy-LORE failed to generate diverse counterfactual examples. Hence, in our future work, we plan to improve the diversity of the generated counterfactual examples. We also plan to extend the current work to provide explanations for multi-class fuzzy-based classifiers.

Acknowledgements

Research projects PI21/00064 and PI18/00169 from ISCIII & FEDER funds and URV grants 2020PFR-B2-61 & 2019PFR-B2-61. First author has a URV Martí Franquès predoctoral grant.

References

- [1] Saleh E, Valls A, Moreno A, Romero-Aroca P, Torra V, Bustince H. Learning fuzzy measures for aggregation in fuzzy rule-based models. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2018. p. 114-27.
- [2] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018;6:52138-60.
- [3] Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019;8(8):832.
- [4] Strumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research. 2010;11:1-18.
- [5] Ribeiro MT, 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. ACM; 2016. p. 1135-44.
- [6] Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:180510820. 2018.
- [7] Maaroof N, Moreno A, Valls A, Jabreel M. Guided-LORE: Improving LORE with a Focused Search of Neighbours. In: Heintz F, Milano M, O'Sullivan B, editors. Trustworthy AI - Integrating Learning, Optimization and Reasoning. Springer; 2021. p. 49-62.
- [8] Maaroof N, Moreno A, Jabreel M, Valls A. Contextualized LORE for Fuzzy Attributes. Artificial Intelligence Research and Development. 2021:435.
- [9] Yuan Y, Shaw MJ. Induction of fuzzy decision trees. Fuzzy Sets and systems. 1995;69(2):125-39.
- [10] Blanco MES, Romero-Aroca P, Pujol RV, Valls A, SaLeh E, Moreno A, et al. A Clinical Decision Support System (CDSS) for diabetic retinopathy screening. Creating a clinical support application. Investigative Ophthalmology & Visual Science. 2020;61(7):3308-8.
- [11] Kommiya Mothilal R, Sharma A, Tan C. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. arXiv e-prints. 2019:arXiv-1905.

Testing Reinforcement Learning Explainability Methods in a Multi-Agent Cooperative Environment

Marc DOMÈNECH I VILA ^a and Dmitry GNATYSHAK ^b and Adrián TORMOS ^b and Sergio ALVAREZ-NAPAGAO ^b

^a Universitat Politècnica de Catalunya, Barcelona, Spain

^b Barcelona Supercomputing Center, Barcelona, Spain

Abstract. The adoption of algorithms based on Artificial Intelligence (AI) has been rapidly increasing during the last years. However, some aspects of AI techniques are under heavy scrutiny. For instance, in many cases, it is not clear whether the decisions of an algorithm are well-informed and reliable. Having an answer to these concerns is crucial in many domains, such as those in which humans and intelligent agents must cooperate in a shared environment. In this paper, we introduce an application of an explainability method based on the creation of a Policy Graph (PG) based on discrete predicates that represent and explain a trained agent's behaviour in a multi-agent cooperative environment. We also present a method to measure the similarity between the explanations obtained and the agent's behaviour, by building an agent with a policy based on the PG and comparing the behaviour of the two agents.

Keywords. Explainable AI, Reinforcement Learning, Policy Graphs, Multi-agent Reinforcement Learning, Cooperative Environments

1. Introduction and Motivation

Over the last decade, methods based on machine learning have achieved remarkable performance in many seemingly complex tasks such as image processing and generation, speech recognition or natural language processing. It is reasonable to assume that the range of potential applications will keep growing bigger in future years. However, there are still many concerns about the transparency, understandability and trustworthiness [1] of systems built using these methods, especially when they are based on so-called *blackbox* models. For example, there is still a need for proper explanations of the behaviour of autonomous vehicles and this could be a risk for their real-world applicability and regulation [2].

Since AI has an increasing impact on people's everyday lives, it becomes urgent to keep progressing on the field of Explainable Artificial Intelligence (XAI) [3]. In fact, there are already regulations that require AI model creators to enable mechanisms that can produce explanations for them, such as the European Union's General Data Protection Regulation (GDPR) that went into effect on

May 25, 2018 [4]. This law creates a “Right to Explanation” whereby a user can ask for the explanation of an algorithmic decision that was made about them. Therefore, XAI is not only desirable but also a frequent requirement. This is also the case for virtual or physical agents trained via Reinforcement Learning (RL), and explainability in RL (XRL) is starting to gain momentum as a different field of XAI.

This paper aims to continue the line of research opened in [5; 6], which consists in producing explanations from predicate-based Policy Graphs (PG) generated from the observation of RL-trained agents in the Cartpole environment. In this paper, we present an application of the same methodology in order to generate explanations for agents trained in a cooperative environment using Multi-Agent Reinforcement Learning (MARL) methods. In the physical world, cooperation between humans and AIs will gradually become more common [7], and thus we believe that it is crucial to be able to explain the behavior of cooperative agents so that their actions are understandable and can be trusted by humans.

Currently, there are several approaches to explain RL agents. In this work, we briefly overview some of them in **Section 2**, we choose one that builds a graph that represents the agent’s behaviour and we apply it to a MARL environment in **Section 4**, also giving insight in how to generate explanations. Once we apply the method, we build a new agent using the graph as a policy in order to compare both agents in **Section 5** and finally, we end with a summary of the main conclusions and contributions from the work done in **Section 6**.

2. Related work

The area of explainability in reinforcement learning is still relatively new, especially when dealing with policies as blackbox models. In this section, we will provide a brief overview of some state-of-the-art XRL methods as well as discuss in more depth the method chosen in this work. A more detailed study of the explainability methods in RL can be found in [8].

According to [8], XRL methods can be classified by their scope of explanation (global/local), timing of explanation (post-hoc/intrinsic), time horizon of explanation (reactive/proactive), type of the environment (deterministic/stochastic), type of policy (deterministic/stochastic) and their agent cardinality (single-agent/multi-agent).

Reactive explanations are those that are focused on the immediate moment. A family of reactive methods is policy simplification, which finds solutions based on tree structures. In these, the agent answers the questions from the root to the bottom of the tree in order to decide which action to take. For instance, Coppens et al. [9] use Soft Decision Trees (SFT), structures that work similarly to binary trees but where each decision node works as a single perceptron that returns, for a given input x , the probability of going right or left. This allows the model to learn a hierarchy of filters in its decision nodes. Another family is reward decomposition, which tries to decompose the reward into meaningful components. In [10], Juozapaitis et al. decompose the Q-function into reward types to try to explain why an action is preferred over another. With this, they can know whether

the agent is taking an action to be closer to the objective or to avoid penalties. Another approach is feature contribution and visual methods, like LIME [11] or SHAP [12], which try to find which of the model features are the most relevant in order to make decisions. On the other hand, Greydanus et al. differentiate between gradient-based and perturbation-based saliency methods [13]. The first ones try to answer the question “Which features are the most relevant to decide the output?” while the latter are based on the idea of perturbing the input of the model in order to analyse how its predictions changes.

Proactive models are those that focus on longer-term consequences. One possible approach is to analyse the relationships between variables. This family of techniques give explanations that are very close to humans because we see the world through a causal lens [14]. According to [15], the causal model tries to describe the world using random variables. Each of these variables has a causal influence on the others. This influence is modelled through a set of structural equations. Madumal et al. generate explanations of behaviour based on a counterfactual analysis of the structural causal model that is learned during RL [16]. Another approach tries to break down one task into multiple subtasks in order to represent different abstraction levels [17]. Therefore, each task can only be carried out if its predecessor tasks have been finished.

According to [18], in order to achieve interoperability, it is important that the tasks had been described by humans. For instance, [17] defines two different policies in hierarchical RL, local and global policies. The first one uses atomic actions in order to achieve the sub-objectives while the second one uses the local policies in order to achieve the final goal.

In addition, there is another approach that combines relational learning or inductive logic programming with RL. The idea behind these methods [19] is to represent states, actions and policies using first order (or relational) language. Thanks to this, it is easier to generalize over goals, states and actions, exploiting knowledge learnt during an earlier learning phase. Finally, another approach consists in building a Markov Decision Process and follow the graph from the input state to the main reward state [16]. This allows us to ask simple questions about the chosen actions. As an optional step, we can simplify the state representation (discretizing it if needed). This step becomes crucial when we are talking about more complex environments [5]. In this work, we will use this last approach. In our case, we will use a method that consists of building a policy graph by mapping the original state to a set of predicates (discretization step) and then repeatedly running the agent policy, recording its interactions with the environment. This graph of states and actions can then be used for answering simple questions about the agent’s execution which is shown at the end of **Section 4**. This is a post-hoc and proactive method, with a global scope of explanation, which works with both stochastic environments and policies and has until now only been tested in single-agent environments.

3. Training agents in a cooperative environment: Overcooked-AI

In this paper, we have used the PantheonRL [20] package for training and testing an agent in Overcooked-AI[21]. Overcooked-AI is a benchmark environment for

fully cooperative human-AI task performance, based on the popular video game **Overcooked**. The goal of the game is to deliver soups as fast as possible. Each soup requires placing up to 3 ingredients in a pot, waiting for the soup to cook, and then having an agent pick up the soup and delivering it. The agents should split up tasks on the fly and coordinate effectively in order to achieve high rewards.

The environment has the following reward function: 3 points if the agent places an onion in a pot or if takes a dish, and 5 points if it takes a soup. Here in this work, we have worked with five different layouts: *simple*, *unident_s*, *random1*, *random0*, and *random3* (**Figure 1**).



Figure 1. Overcooked layouts (*simple*, *unident_s*, *random1*, *random0*, *random3*)

At each timestep, the environment returns a list with the objects not owned by the agent present in the layout and, for each player, the position, orientation, and object that it is holding and its information. We can also get the location of the basic objects (dispensers, etc.) at the start of the game.

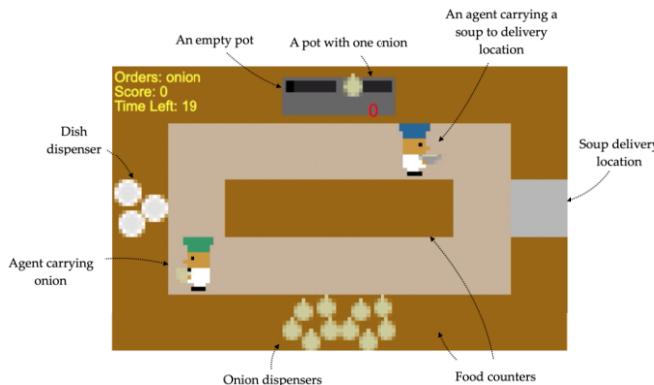


Figure 2. Overcooked-AI game dynamics.

For example, the agent would receive the following data from **Figure 2**:

- **Player 1:** Position (5, 1) - Facing (1, 0) - Holding Soup
- **Player 2:** Position (1, 3) - Facing (-1, 0) - Holding Onion
- **Not owned objects:** Soup at (4, 0) - with 1 onion and 0 cooking time.

The aim of our work is not to solve the Overcooked game but rather to analyze the potential of explainability in this cooperative setting. Therefore, we do not really care about what method is used to train our agent. However, it is important that the agent performs reasonably well in order to verify that we are explaining an agent with a reasonable policy. Therefore, we have used the Proximal Policy

Optimization (PPO) algorithm because it is one of the methods that has achieved the best results in [21]. Indeed, if we get good results with PPO, we should also get good results with other methods since this explainability method is independent from the RL method used. In our case, we have trained five different agents (one for each layout) for 1M total timesteps and with an episode length of 400 steps. The results are the following:

- *simple*: Mean Episode Reward = 387.87 and Standard Deviation = 25.33
- *unident_s*: Mean Episode Reward = 757.71 and Standard Deviation = 53.03
- *random0*: Mean Episode Reward = 395.01 and Standard Deviation = 54.43
- *random1*: Mean Episode Reward = 266.01 and Standard Deviation = 48.11
- *random3*: Mean Episode Reward = 62.5 and Standard Deviation = 5.00

4. Building a policy graph for the trained agent

As we mentioned in **Section 2**, we have chosen a method based on the creation of a policy graph. This method consists of building a directed graph where each node represents a state, and each edge represents the transition going from one node to another taking a specific action.

In order to build the policy graph, we need to discretise the state representation. This step is crucial since we need to map each state to a node in our PG. We have created a total of 10 predicates to represent each state.

The first two are *held* and *held_partner*, which have 5 possible values depending on which object the agent and its partner, respectively, are holding (e.g., "O" for "Onion" object or "*" if is not holding anything). The third is *pot_state*, which has 4 possible values depending on the state of each pot:

- "Of", when [pot.onions = 0].
- "Fi", when [pot.onions = 3 ∧ pot.timer = 20].
- "Co", when [pot.onions = 3 ∧ pot.timer < 20].
- "Wa", when [pot.onions < 3].

To relate the actions of the agent with the relative position of the objects, we introduce another 6 predicates: *onion_pos(X)*, *tomato_pos(X)*, *dish_pos(X)*, *pot_pos(X)*, *service_pos(X)* and *soup_pos(X)*. All of them with the same 6 possible values depending on the next action to perform to reach the object quickly (e.g., "T" for "Top" action). The last one is *partner_zone*, intended to help the agents cooperate along with *held_partner*. It has 8 possible values depending on which cardinal point the partner is located (e.g., "NE" for "North East").

As we mentioned in **Section 2**, the aim of the PG algorithm is to apply the same method as in **Figure 3**: record all the interactions of the original trained agent by executing it in a large set of random environments and build a graph relating predicate-based states and actions. Our work has followed two approaches for building this graph:

- **Partial Policy Graph**: This algorithm builds a directed graph. For each state, it takes the most probable action. Therefore, we do not add all the agent interactions to our graph, only those that belong to the most used

action by the agent. This means that for each node, we only have one possible action (the most probable in this case). We think this could be an interesting approach since we are building a deterministic agent.

- **Complete Policy Graph:** This algorithm builds a multi-directed graph. For each state, it takes all the agent interactions, adding them to our graph. This means that for each node, we have multiple possible actions with an associated probability. We think this could be an interesting approach since we are maintaining stochasticity.

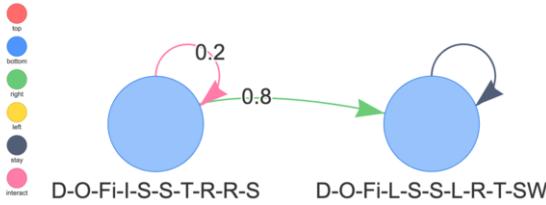


Figure 3. Extract of two states from a Complete PG generated from Overcooked.

Once we have built our PG (regardless of the PG algorithm), we have access to the generation of natural language explanations as in [5], where the authors validate the PG by comparing the sentences generated by the algorithm against sentences written by human experts. While this is valid as a qualitative approach for validation, it becomes obvious that this method heavily depends on experts and on the nature of the specific domain. However, the authors propose three questions to help explain the agent’s behaviour. These questions are:

1. **What will you do when you are in state X ?**: Look for the possible actions in the policy graph from the input state that the user wants to check.
2. **When do you perform action X ?**: Look for these states where the action X is the most probable action.
3. **Why did you not perform action X in state Y ?**: Look at which state it would reach if it had chosen action X in state Y . Once we have those nearby (similar) states, then we compute the difference between both states.

5. Validating the policy graph

In this section, we study how to validate our Policy Graph. To do it, we build another agent that follows a policy based on the PG we obtained as explained in **Section 4**: at each step, the agent receives the current state and decides its next action by querying the PG for the most probable action from the most similar state to the current one. In order to test this new agent, we run it multiple times in random environments. If we discretize the space as explained in **Section ??**, we have 10 predicates and 37,324,800 potential states. This means it is very likely that the new agent will find states never seen before, so it is very important to set up a strategy to deal with these situations. We introduce a state similarity metric to deal with unknown states: we consider that two states (S_1, S_2) are similar when

$diff(S_1, S_2) \leq 1$, where $diff(S_1, S_2)$ computes the number of different predicate values one each other. For example, if we have the states $S_1 = O\text{-Co-S}$ and $S_2 = O\text{-Co-N}$, then $diff(S_1, S_2) = 1$ so they are considered similar.

As we saw in **Section 4**, we are testing 2 different algorithms. Therefore, we have built 2 new agents for each layout. Now, we will see how each of these algorithms make decisions. Assuming we are in state S , we can distinguish 3 cases:

1. $S \in PG$: Picks an action using weights from the probability distribution in the PG.
2. $S \notin PG$, but a similar state is found: Same as case 1 but using the similar state.
3. $S \notin PG$ and a similar state is not found: Pick a random action.

All the agents have been trained using batches of 25 seeds. Altogether, the training consisted in 500 seeds and 3 episodes per seed. In order to validate the new agent, we have generated 3 metrics: Transferred Learning (TL), the ratio between the RL agent average episode reward and the new agent's average episode reward; Standard Deviation (STD), the ratio between the RL agent average standard deviation of reward and the new agent's standard deviation of reward; and New States (NS) visited, the proportion between previously unknown and total visited states. In order to do comparisons, we have tested both PG algorithms using multiple discretizers (D11, D12, D13 and D14). D11 has the predicates *held*, *pot_state* and *predicate_pos*, D12 has D11 predicates plus *held_partner*, D13 has D11 predicates plus *partner_zone* and D14 has all predicates.

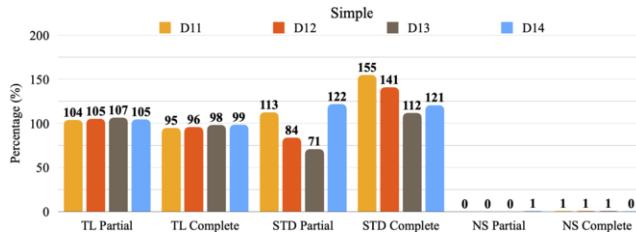


Figure 4. Results from layout *simple*

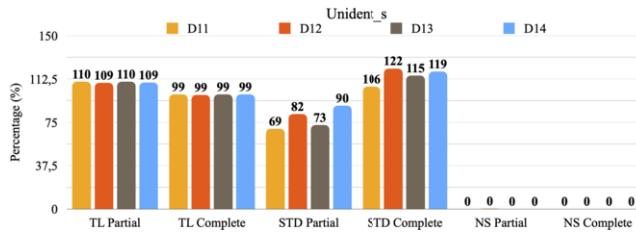


Figure 5. Results from layout *unident_s*

Figures [4-8] show the results we have obtained from the different agents. We can see that in the *simple* and *unident_s* scenarios, the Partial agents manage to outperform the original ones while the Complete agents do not. On the other hand,

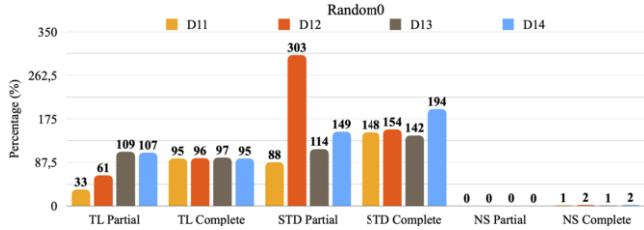


Figure 6. Results from layout random0

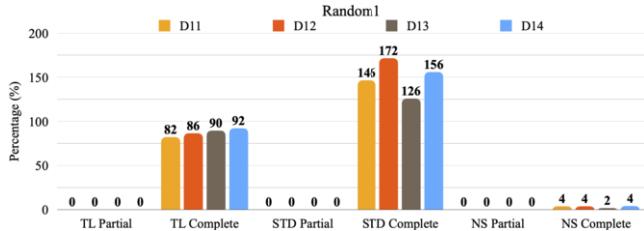


Figure 7. Results from layout random1

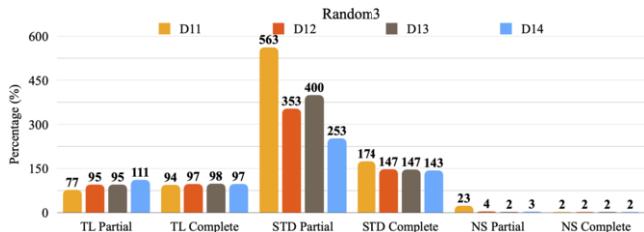


Figure 8. Results from layout random3

it seems that in these cases it is not necessary to introduce cooperative predicates to explain their behaviour, unlike in the *random0* and *random3* scenarios. For instance, in *random0* the Partial agent needs the predicate *partner_zone* and in *random3* the predicate *held_partner* to get good results. We can also see how in *random1*, the Partial agent is not even able to score even though the Complete algorithm scores quite well. Regarding the standard deviation of these agents, it is usually higher than the original agent's, likely because the PG we have built is based on the simplification of states and actions, so the policy also ends up being a simplification. Finally, we can see how the vast majority of agents achieve a really low NS percentage, which means that the agent knows exactly what action to take most of the time.

To summarise, according to Figures [4-8], the Complete algorithm is more stable in our 3 metrics than the Partial one, as the latter can score zero points or perform better than the original agent depending on the layout, probably due to its deterministic nature. Therefore, the Complete algorithm is the more reliable from an XRL point of view and should be the one used for producing explanations.

On the other hand, we have also seen that although there are scenarios where it is not necessary to introduce cooperative predicates to explain the agent's

behaviour, there are others where this information is crucial, which makes sense due to the fact that the layout influences the need for cooperation.

6. Conclusions

XAI is a research area that is growing by leaps and bounds in recent years, due to the need to understand and justify the decisions made by AIs, especially in the field of RL. All the research in this area can be key not only to study the quality of an agent's decision but also to help people rely on AI, especially in situations where humans and machines have to cooperate, and it is becoming necessary to be able to give explanations about their decisions. There are already some proposals in the literature to provide them, and it is important to test their effectiveness in practice.

In this work, we have used an explainability method based on the construction of a PG by discretising the state representation into predicates for later applying it to a cooperative MARL environment (Overcooked). We have proposed two different algorithms to generate the PG and we have managed to give some explanations that according to [5] should be validated by human experts with domain-specific knowledge. In order to validate the PG, we have applied a method already tested in [6] to generate automatically policies based on these explanations to build agents that represent them. Finally, we have seen that this technique is capable of giving explanations in a MARL cooperative environment like Overcooked.

7. Acknowledgement

This work has been partially supported by the H2020 knowlEdge European project (Grant agreement ID: 957331).

References

- [1] Li BO, Qi P, Liu BO, Di S, Liu J, Pei J, et al. Trustworthy AI: From Principles to Practices. oct.
- [2] Omeiza D, Webb H, Jirotka M, Kunze L. Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*. 2021;1-21. ArXiv:2103.05154 [cs].
- [3] Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A. Explainable artificial intelligence: Concepts, applications, research challenges and visions. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer; 2020. p. 1-16.
- [4] Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*. 2017;38(3):50-7.
- [5] Hayes B, Shah JA. Improving robot controller transparency through autonomous policy explanation. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI. IEEE; 2017. p. 303-12.

- [6] Climent A, Gnatyshak D, Alvarez-Napagao S. Applying and Verifying an Explainability Method Based on Policy Graphs in the Context of Reinforcement Learning. In: Artificial Intelligence Research and Development. IOS Press; 2021. p. 455-64.
- [7] Dafoe A, Hughes E, Bachrach Y, Collins T, McKee KR, Leibo JZ, et al. Open Problems in Cooperative AI. arXiv:201208630 [cs]. 2020 Dec. ArXiv: 2012.08630.
- [8] Krajna A, Brcic M, Lipic T, Oncevic J. Explainability in reinforcement learning: perspective and position. arXiv preprint arXiv:220311547. 2022.
- [9] Coppens Y, Efthymiadis K, Lenaerts T, Nowé A, Miller T, Weber R, et al. Distilling deep reinforcement learning policies in soft decision trees. In: Proceedings of the IJCAI 2019 workshop on explainable artificial intelligence; 2019. p. 1-6.
- [10] Juozapaitis Z, Koul A, Fern A, Erwig M, Doshi-Velez F. Explainable reinforcement learning via reward decomposition. In: IJCAI/ECAI Workshop on explainable artificial intelligence; 2019. .
- [11] Ribeiro MT, 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, San Francisco, CA, USA, August 13-17, 2016; 2016. p. 1135-44.
- [12] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765-74.
- [13] Greydanus S, Koul A, Dodge J, Fern A. Visualizing and understanding atari agents. In: International conference on machine learning. PMLR; 2018. p. 1792-801.
- [14] Sloman S. Causal models: How people think about the world and its alternatives. Oxford University Press; 2005.
- [15] Halpern JY, Pearl J. Causes and Explanations: A Structural-Model Approach — Part 1: Causes. 2013 jan.
- [16] Madumal P, Miller T, Sonenberg L, Vetere F. Explainable reinforcement learning through a causal lens. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34; 2020. p. 2493-500. Issue: 03.
- [17] Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in neural information processing systems. 2016;29.
- [18] Shu T, Xiong C, Socher R. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. arXiv preprint arXiv:171207294. 2017.
- [19] Zambaldi V, Raposo D, Santoro A, Bapst V, Li Y, Babuschkin I, et al.. Relational Deep Reinforcement Learning. arXiv; 2018. Number: arXiv:1806.01830 arXiv:1806.01830.
- [20] Sarkar B, Talati A, Shih A, Sadigh D. PantheonRL: A MARL Library for Dynamic Training Interactions. arXiv; 2021. Number: arXiv:2112.07013 arXiv:2112.07013 [cs]. Available from: <http://arxiv.org/abs/2112.07013>.
- [21] Carroll M, Shah R, Ho MK, Griffiths T, Seshia S, Abbeel P, et al. On the utility of learning about humans for human-ai coordination. Advances in neural information processing systems. 2019;32.

Mutual Information Weighing for Probabilistic Movement Primitives

Adrià COLOMÉ and Carme TORRAS

Institut de Robòtica i Informàtica Industrial, CSIC-UPC

Abstract. Reinforcement Learning (RL) of trajectory data has been used in several fields, and it is of relevance in robot motion learning, in which sampled trajectories are run and their outcome is evaluated with a reward value. The responsibility on the performance of a task can be associated to the trajectory as a whole, or distributed throughout its points (timesteps). In this work, we present a novel method for attributing the responsibility of the rewards to each timestep separately by using Mutual Information (MI) to bias the model fitting of a trajectory.

Keywords. Reinforcement Learning, Motion Learning, Mutual Information

1. Introduction

In the application of AI to robot motion learning, it is a common approach to use a policy represented as a set of parameters that, applied to a parametrized family of functions, define the robot motion at every timestep. These policy parameters are then perturbed in order to generate new motions that are evaluated by a reward function. Then, a so-called Policy Search (PS) [2,3] method obtains a new policy (set of parameters) from the samples and their measured performance. In generating these new samples, two approaches were defined in [2]: step-based and episodic-based.

In step-based sampling, a perturbation of the policy is generated at every timestep of the trajectory and so each trajectory is given a score (reward) value at every timestep. On the contrary, episode-based sampling perturbs the parameters of a whole trajectory and runs it. Whilst the step-based approach can provide more information when timestep performance can be measured, its application to fields such as robotics can become dangerous, as it induces high-frequency perturbations on the reference motion. This is why, in robotics, step-based methods have been applied to episode-based samples, keeping the idea of associating a reward to each timestep of a sampled trajectory, but with a smoother profile. As an example of such methods, in Policy Improvement with Path Integrals (PI2) [4], the authors define a cost-to-go associated to each sample in order to accordingly give importance to each timestep of each trajectory. This cost-to-go is then associated with a probability used as a weight for updating the policy parameters. This method assigns a responsibility to each timestep by using the time remaining for the trajectory and its deviation from the mean. However, there is no necessary correlation between the deviation from the mean and the reward. A trajectory could present a lot of variability that is completely unrelated to its performance. Therefore, in this work, we explore a way to distribute responsibility by using an indicator of how two variables (deviation from the

mean at current timestep vs whole trajectory reward) are related, namely Mutual Information (MI) [1]. By using MI, we obtain relative weights with which to bias the fitting of the updated policy, i.e., the trajectory stochastic parametrization. In this paper, we firstly present preliminary concepts in Sec. 2, followed by the proposed methodology in Sec. 3 and the conclusions in Sec. 4.

2. Preliminaries

2.1. Mutual Information

In this work, we will use the MI between two variables and its discretization. In general, the mutual information between two random variables is a measure of their dependence, or how much information is obtained from one variable after observing the other, and is calculated as $MI(X;Y) = \sum_Y \sum_X P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$. In our applications, we neither have discrete distributions nor have such analytical expressions so as to calculate the MI. Therefore, we partition the data in bins of an equal number of data points and the discrete version is used. Using such bins will prevent the method from outliers and nonlinearities.

2.2. Probabilistic Movement Primitives (ProMPs)

For motion characterization, Movement Primitives (MP) are a popular approach in order to easily encode a trajectory. Amongst the most used MPs, Probabilistic Movement Primitives (ProMPs) [5] present advantages such as that they can fit the time-dependent variability of the motion, and probability operations such as product and conditioning can be used on them. Given a number of basis functions per DoF, N_f , ProMPs use a time-dependent matrix $\Phi_t = [\phi_t^1; \dots; \phi_t^{N_f}]$ to encode position, ϕ_t being the vector of normalized kernel basis functions (e.g., uniformly distributed Gaussian basis function over time). Thus, the position and velocity state vector y_t can be represented as $y_t = \Phi_t^T \omega + \varepsilon_y$, where $\varepsilon_y \sim \mathcal{N}(0, \Sigma_y)$ is a zero-mean Gaussian noise and the weights ω are also treated as random variables with a distribution $p(\omega) = \mathcal{N}(\omega | \mu_\omega, \Sigma_\omega)$. Subsequently, the parameters of the distribution $\theta = \{\mu_\omega, \Sigma_\omega, \Sigma_y\}$, Σ_y being the state covariance, are fitted by means of a maximum likelihood estimate, i.e., we compute the sample mean and the sample covariance of ω . Then the probability of observing a trajectory τ can be expressed as the product of all timestep probabilities: $p(\tau; \theta) = \prod_t \int \mathcal{N}(y_t | \Phi_t^T \omega, \Sigma_y) \mathcal{N}(\omega | \mu_\omega, \Sigma_\omega) d\omega$.

3. Methodology

We assume we have a set of N_k trajectories with dimension D , and composed on N_t timesteps each. Therefore, y_{td}^k will correspond to the d -dimension on the t -th timestep of the k -th trajectory. Associated with each trajectory, we have a reward R_k that is a performance indicator. For each timestep t , we will compute the mutual information $MI(\{y_{td}^k\}_{k \in K}, \{R^k\}_{k \in K})$. Then, we can use an Expectation Maximization-based approach with ProMPs, such as in [6]. However, in [6] the authors considered that the reward for each trajectory has to be associated to the trajectory as a whole. In this paper, we ap-

ply the same method with the following differences: The previous reference use a linear dimensionality reduction in order to reduce the state space's dimension. Here, we omit this part, despite it would be straight-forward to add it if necessary. In this paper, we compute the mutual information between the rewards of the demonstrations and the variability of trajectories and use it to weigh data. Additionally, in order to penalize higher mutual information values in low variance parts of the trajectory, we use the term $\eta_t = \text{trace}(\Sigma_y) \cdot \text{MI}(\{R_k\}_{k=1..N_k}, \{y_{tk}\}_{k=1..N_k})$. By adding a factor of the trace of the variance at each time-step, we enforce that the term η_t is higher when the trajectory values at that timestep present a high correlation with the reward outcome, in addition to these trajectory values also presenting a higher variability.

As we also know the rewards R_k (performance of each trajectory), we used Relative Entropy Policy Search (REPS) in order to convert them to trajectory weights d_k but, differently from [6], we assign each data point y_{tk} an importance equal to $\delta_{k,t} = \eta_t \cdot d_k$.

3.1. Expectation-Maximization

In this section, we will point out to the main differences with [6] in the formulation. The Expectation Maximization method consists of two steps: First, we evaluate the probability of the model parameters given each trajectory \mathbf{Y}_k (vectorized as a column vector), $p(\mathbf{Y}^k|\omega) = \mathcal{N}(\mathbf{Y}^k|\Psi^T, \mathbf{I}_{N_t} \otimes \Sigma_y)$, where \otimes is the kronecker product and Ψ^T is the matrix of concatenated kernel functions ϕ_t for different dimensions, i.e., $\Psi^T = [I_d \otimes \Phi_1; \dots; I_d \otimes \Phi_{N_t}]$. By operating and using Bayes' rule, we can obtain that $p(\omega|\mathbf{Y}^k) = \mathcal{N}(\mu_k, \Sigma_k)$, with

$$\begin{aligned}\mu_k &= \mu_\omega + \Sigma_\omega \Psi (\mathbf{I}_{N_t} \otimes \Sigma_y + \Psi^T \Sigma_\omega \Psi)^{-1} (\mathbf{Y}^k - \Psi^T \mu_\omega) \\ \Sigma_k &= \Sigma_\omega - \Sigma_\omega \Psi (\mathbf{I}_{N_t} \otimes \Sigma_y + \Psi^T \Sigma_\omega \Psi)^{-1} \Psi^T \Sigma_\omega.\end{aligned}$$

We will then compute these μ_k, Σ_k , and use them on the Maximization step, where we maximize the weighed expectation of the log-likelihood function in order to obtain the new parameters θ : $L = \sum_{k=1}^{N_k} \sum_{t=1}^{N_y} \delta_{tk} \mathbb{E}_{\omega|y_t^k; \theta^{old}} [\log(p(\omega, y_t^k; \theta))]$. By solving this equation (see [6]), we obtain that, at each iteration, the new policy parameters are:

$$\begin{aligned}\mu_\omega &= \left(\sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} \right)^{-1} \sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} \mu_k \\ \Sigma_\omega &= \left(\sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} \right)^{-1} \sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} \left[\Sigma_k + (\mu_\omega - \mu_k)(\mu_\omega - \mu_k)^T \right], \\ \Sigma_y &= \left(\sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} \right)^{-1} \sum_{k=1}^{N_k} \sum_{t=1}^{N_t} \delta_{tk} [\Psi^T \Sigma_k \Psi^T + (\mathbf{y}_t^k - \Psi^T \mu_k)(\mathbf{y}_t^k - \Psi^T \mu_k)^T].\end{aligned}$$

3.2. Results

In order to test this approach, we generated a set of random time-dependent trajectories with high variance around a viapoint and on the endpoint (see Fig. 1 left with the trajectories and their respective mean and variance as a shaded area), and low variance elsewhere. We defined a reward function for each trajectory that penalizes its distance from the desired via-point (in red), i.e., the reward is determined by how far the trajectory passes by a certain via-point. Then, we applied the method in [6] with $r = d$ in the paper, and our method by using REPS [7] to convert trajectory weights to rewards. The

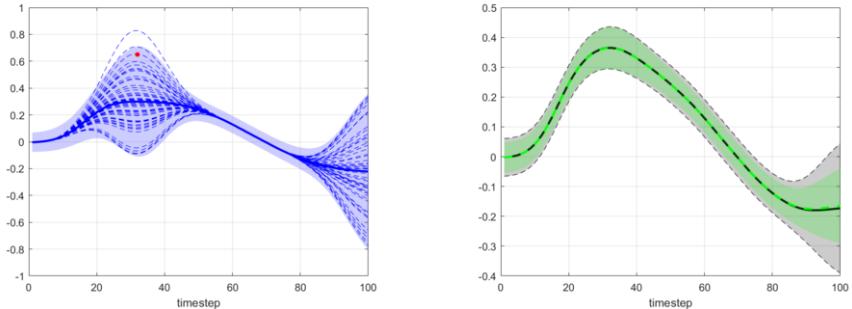


Figure 1. Left: data trajectories used for the proof of concept experiment and their fitting with least-squares. The red dot marks the desired via-point and the reward associated to each trajectory is the distance from that viapoint. On the right plot, the proposed method (in black) against the reference [6]. We see that, while both methods provide a similar result in the area where the reward is decided, at the end of the trajectory, the proposed method allows for more variance, given the fact that the end position does not affect the reward. Note different scales in the vertical axis.

results of fitting both methods are shown in Fig. 1 right, where the mean of both methods (ours in black, [6] in green) is similar, but the variance at the endpoint is larger in our method. This is mostly because the MI between the reward values and the endpoint of the trajectories is small, therefore our method requires less precision and does not give importance on the final point.

4. Conclusions

In this ongoing work paper, we devised a methodology for re-weighing data in order to better distribute the responsibility of the reward throughout a trajectory. The preliminary results in a proof of concept task displayed in Fig. 1 how the method is relieving of responsibility those parts of the trajectory that do not provide a gain to the reward, and therefore allow for more variance. In other words, relating the importance of a trajectory data-point to how much it correlates with the overall performance at that point in the trajectory allows for a better fitting of the trajectory.

References

- [1] Cover TM, Thomas JA. Entropy, relative entropy and mutual information. In: Elements of Information Theory. John Wiley & Sons, Inc. ISBN 9780471241959, 2001.
- [2] Deisenroth MP, Neumann G, Peters G. A Survey on Policy Search for Robotics. Foundations and Trends in Robotics. 2(1-2): 1-141, 2013.
- [3] Chatzilygeroudis K, Vassiliades V, Stulp F, Calinon S, Mouret JB. A survey on policy search algorithms for learning robot controllers in a handful of trials. IEEE Trans. on Robotics. 36(2): 328-347, 2020.
- [4] Stulp F, Theodorou EA, Schaal S. Reinforcement Learning With Sequences of Motion Primitives for Robust Manipulation. IEEE Transactions on Robotics 28(6):1360-1370, 2012.
- [5] A. Paraschos, G. Neumann, C. Daniel, and J. Peters, "Probabilistic movement primitives". In *Advances in NIPS*, Cambridge, MA: MIT Press., 2013.
- [6] A. Colomé, G. Neumann, J. Peters and C. Torras "Dimensionality Reduction for Probabilistic Movement Primitives". *IEEE-RAS Humanoid Robots*, pp. 794-800, 2014.
- [7] J. Peters, K. Mülling and Y. Altün, "Relative Entropy Policy Search". *Twenty-Fourth National Conf. on Artificial Intelligence*, pp 182-189, 2011.

Subject Index

aerospace industry	151	decision support system	105
aggregation function	13, 45	deep learning	59, 67, 225, 229, 243, 249, 259, 269, 308
algorithms	115	deep neural network	105
Anova	67	dendrogram	172
anticipatory shipping	147	diabetic retinopathy	308, 345
Artificial Intelligence	155	digital breast tomosynthesis	269
automotive OEM	147	digital twin(s)	83, 155
autonomous agents	160	digital twins and synthetic data	83
Barcelona geolocation	279	disease anticipation	164
belief merging	7	domain generalization	221
bias	59, 325	ensemble classifiers	181
biomedical imaging	249	evaluation metrics	325
blind image quality	243	explainability	325
bootstrapping	172	Explainable AI (XAI)	3, 345, 355
breast cancer	298	exudates	308
breast cancer classification	269	F1 score	147
bright-field microscopy	249	feature attribution methods	325
Building Information Modeling	95	federated learning	91
building regulations	95	flows	191
CAD	289	Focal Modulation Networks	279
CAD systems	298	focus	325
case-based reasoning	87	food-porn	67
classification	105, 147	fundus images	308
cloth folding dataset	199	fungicide management	164
cluster validity indices	172	Fuzzy Decision Tree	345
clustering	71, 143, 168	Fuzzy Random Forest	181
CNN	71, 239	G-Protein Coupled Receptors	209
combinatorial optimization	17	garment manipulation	199
completeness	35	gender	59
computer vision	269	graph	71
computer-based simulation	160	Graph Convolutional Network	125
context-aware recommendations	125	group recommender systems	115
contingency tables	168	hierarchical clustering	172
conversational agents	335	horn fragment	7
cooperative environments	355	image analysis	249
COVID-19	229	image classification	279
culture	71	image enhancement	239
CURE clustering	172	image segmentation	249
customer delivery time distribution	147	implementation	7
darknet traffic	105	importance-performance analysis	45
data spaces	91	Industry 4.0	151
DBSCAN	71	influence	71

information aggregation	45	polarization	17
instance segmentation	225	policy graphs	355
intellectual and developmental disability	139	postulates for consensus analysis	321
Intelligent Decision Support	87	predictive quality	151
IOWA operator	45	prevent economic losses	164
learning by demonstration	199	priority of care	139
lesion segmentation	308	probabilistic weighted interactions	321
LiDAR	259	proof systems	25
LORE	345	prostate segmentation	221
Lotka-Volterra system	191	pseudo-reference	243
LSTM	209	public media service	125
lung cancer	289	quality of life	139
machine ethics	335	R	143
machine learning	3, 59, 105, 139, 147, 151, 155, 164, 345	radiomics	289, 298
many-valued logics	7	rational function	13
Markov chain Monte Carlo	168	recommender systems	125
Mask-RCNN	225	regression	67
MaxCUT	25	reinforcement learning	335, 355, 365
maximum satisfiability	35	reliability	289
MaxSAT	25	remote sensing	259
medical image processing	229	reproducibility	115
memes	71	resource saving	164
metrics	55	road crack	225
molecular dynamics	209	rule-compliance checking	95
mosaics	325	SAT	25
motion learning	365	scoring functions	143
MRI imaging	221	segmentation	259
multi-agent reinforcement learning	355	semantic tableaux	35
multi-class ordinal classification	181	sentimental analysis	67
multi-objective genetic algorithm	9	signed logic	7
multiagent systems	160	similarity measures	243
mutual information	168, 365	smart city policy	83
networks	143	social media	71
neural networks	191	social network(s)	17, 71
Neural Ordinary Differential Equations	191	stochastic block model	143
NLP	55, 59	submarine images	239
non parametric bootstrap	143	support and attack relations	321
numerical simulation	155	support paradigm	139
object detection	225, 239	support vector machine	269
online data imputation	87	symbolic AI	3
online discussion	321	symmetry	13
ontologies	95	synthetic data	83, 91
optimization	87	text mining	55
outdoor image geolocation	279	text pre-processing	55
OWA operator	181	time-series	87
point clouds	259	Tor	105
		transfer learning	221, 229
		transparency	3
		travel route optimization	9

trustworthy AI	3	video games	335
Twitter	59	virtual reality framework	199
ultrasound imaging	298	vision transformers	298
unseen target	221	web tool	55
user experience questionnaires	335	weighted average objective	
valued arguments	321	balancing	9
vehicle reallocation	147	X-ray imaging	229

This page intentionally left blank

Author Index

Abad, M.	229	Catalán, I.A.	239
Abdel-Nasser, M.	139, 225, 243, 269, 289, 298, 308	Cerquides, J.	59, 249, 279
Abio, A.	155	Coll, J.	35, 95
Agell, N.	45	Coll, X.	95
Aguiló, I.	13	Colomé, A.	365
Ahmed, B.	243	Cortés, A.	v
Alcaina, G.	279	Cortés, U.	83
Ali, M.Y.S.	308	Costa, V.	7
Alimoradi, M.	105	Cozar-Alier, C.	160
Almirall, E.	83	Cruz, S.A.	155
Alsinet, T.	17, 321	Cugueró-Escofet, M.À.	87
Álvarez-Ellacuría, A.	239	Da Silva, M.	155
Álvarez-García, E.	55	Daliri, A.	105
Alvarez-Napagao, S.	355	Dellunde, P.	7
Angel-Velez, D.	249	Domènec i Vila, M.	355
Angulo, C.	91	Domingo, X.	151
Ansótegui, C.	25	Dueñas, S.	139
Arcos, J.-L.	249	Duran, P.G.	125
Argelich, J.	17, 321	Escalera, S.	71
Arias-Duart, A.	325	Etxegarai, M.	151
Arratia, A.	143, 168, 191	Falomir, Z.	67
Athanasiou, G.	249	Fernandez-Marquez, J.L.	279
Azari-Dolatabad, N.	249	Flaminio, T.	v
Baget, M.	308	Foix, S.	199
Basile, B.	164	Fronte, P.	45
Béjar, R.	17, 321	Garcia, P.	95
Ben Loussaief, E.	221	García-Costa, D.	55
Boix-Granell, A.	199	Garcia-Gasulla, D.	325
Bonada, F.	151, 155	García Sánchez, J.M.	147
Boratto, L.	115	Giannotti, F.	3
Borras, J.	9	Gibert, K.	172
Brígido, A.	155	Giménez-Ábalos, V.	325
Brugarolas, D.	45	Gnatyshak, D.	355
Bruins, P.	83	Gombert, A.	59
Busqué, R.	155	Gonzalez-Abril, L.	67
Callegaro, D.	83	Grimaldo, F.	v, 55
Camps, M.	151	Habet, D.	35
Camps-Ortin, I.	160	Hassan, L.	269
Carós, M.	259	Hassanien, M.A.	298
Casacuberta, C.	71	Huitzil, I.	95
Casales-Garcia, V.	67	Jabreel, M.	308, 345
Casas-Roma, J.	229	Just, A.	259
		Karna, A.	172

König, C.	209	Pujol, A.J.	164
Kumar Singh, V.	298	Raes, A.	249
Lacheny, W.	151	Rashed, A.	243
Lerma Martín, A.	147	Rashwan, H.A.	139, 225
Levy, J.	25	Raya, C.	91
Lhotska, L.	181	Renedo-Mirambell, M.	143, 168
Li, C.-M.	35	Riera, J.V.	13
Li, S.	35	Rodríguez, I.	335
Lisani, J.-L.	239	Rodríguez-Aguilar, J.A.	335
Llido, D.M.	67	Rodríguez-Soto, M.	335
López-Correa, J.M.	209	Romaní, M.	191
Lopez-Sánchez, M.	335	Roselló-Marín, E.	335
Maaroof, N.	345	Sabater-Mir, J.	160
Manyà, F.	35	Sadeghi, R.	105
Martínez, P.	83	Salamó, M.	115
Martínez, S.	17	Saleh, A.	269
Massanet, S.	13	Sànchez-Marrè, M.	87
Moreno, A.	9, 181, 345	Santamaría, M.	83
Mülâyim, M.O.	279	Sanz, I.	67
Murgese, F.	279	Sbert, C.	239
Musceros, L.	67	Schorlemmer, M.	95
Nandi, G.C.	139	Seguí, S.	259
Okran, A.M.	225	Silveira, J.D.	115
Omer, O.A.	243	Sledzik, R.	105
Onielfa, C.	71	Tahmooresi, M.	289
Onrubia, J.R.	164	Tormos, A.	355
Orama, J.A.	9	Torras, C.	199, 365
Ortiz, C.	191	Torrens, M.	45
Osman, N.	95	Torres, P.	155
Otero, M.	164	Valls, A.	181, 308, 345
Palmer, M.	239	Van Soom, A.	249
Parés, F.	325	Velasquez, L.F.	164
Pascual-Fontanilles, J.	181	Vellido, A.	209
Pascual-Pañach, J.	87	Vidales, B.M.	139
Pauleau, S.	151	Vilasís Cardona, X.	147
Petro, A.B.	239	Vitrià, J.	125, 259
Pijuan, J.	164	Yadav, G.K.	139
Prados, F.	229	Zabihimayyan, M.	105
Puig, D.	139, 221, 225, 243, 269, 289, 298		